The American Economic Review

5/29

Cuk-H/02172-104-P0110

ARTICLES

104

ROBERT C. MERTON
Applications of Option-Pricing Theory: Twenty-Five Years Later

MYRON S. SCHOLES
Derivatives in a Dynamic Environment

HAO LI AND SHERWIN ROSEN
Unraveling in Matching Markets

W. BENTLEY MacLEOD AND JAMES M. MALCOMSON
Motivation and Markets

KAUSHIK BASU AND PHAM HOANG VAN
The Economics of Child Labor

WULONG GU AND PETER KUHN
A Theory of Holdouts in Wage Bargaining

JOSHUA D. ANGRIST AND WILLIAM N. EVANS
Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size

DONALD R. DAVIS Does European Unemployment Prop Up American Wages? National Labor Markets and Global Trade

GEORGE W. EVANS, SEPPO HONKAPOHJA, AND PAUL ROMER
Growth Cycles

COSTAS AZARIADIS AND BRUCE SMITH
Financial Intermediation and Regime Switching in Business Cycles

ROSS LEVINE AND SARA ZERVOS
Stock Markets, Banks, and Economic Growth

RAGHURAM G. RAJAN AND LUIGI ZINGALES
Financial Dependence and Growth

STEPHEN MORRIS AND HYUN SONG SHIN
Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks

SHORTER PAPERS: J. Hersch; L. Friedberg; J. S. Gray; Y.-K. Che and I. L. Gale; P. Pecorino.

JUNE 1998

P-11027

THE AMERICAN ECONOMIC ASSOCIAT

• Printed at Banta Company, Menasha, Wisconsin, U.S.A.

• Copyright © 1998 by the American Economic Association. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation, including the name of the author. Copyrights for components of this work owned by others than AEA must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works, requires prior specific permission and/or a fee. Permissions may be requested from the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203.

• No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, back issues, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

A032901

THE AMERICAN ECONOMIC REVIEW (ISSN 0002-8282), June 1998, Vol. 88, No. 3, is published five times a year (March, May, June, September, December) and every four years it is published six times a year (twice in December) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: $55.00, $66.00, or $77.00, depending on income. A membership also includes subscriptions to *The Journal of Economic Literature* and *The Journal of Economic Perspectives*. In countries other than the U.S.A., add $23.00 for extra postage. Information on becoming a member can be found on the last page of this journal. Periodical postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the

B. C. ROY MEMORIAL LIBRARY
INDIAN INSTITUTE OF MANAGEMENT CALCUT

Founded in 1885

Officers

President
ROBERT W. FOGEL
University of Chicago

President-elect
D. GALE JOHNSON
University of Chicago

Vice-Presidents
ROBERT J. BARRO
Harvard University
JUNE E. O'NEILL
Congressional Budget Office

Secretary
JOHN J. SIEGFRIED
Vanderbilt University

Treasurer
C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review
ORLEY ASHENFELTER
Princeton University

Editor of The Journal of Economic Literatu
JOHN McMILLAN
University of California-San Diego

Editor of The Journal of Economic Perspec
ALAN B. KRUEGER
Princeton University

Executive Committee

Elected Members of the Executive Committe
RONALD G. EHRENBERG
Cornell University
BARBARA L. WOLFE
University of Wisconsin
RACHEL McCULLOCH
Brandeis University
PAUL M. ROMER
Stanford University
ANGUS S. DEATON
Princeton University
LAURENCE J. KOTLIKOFF
Boston University

EX OFFICIO Members
ANNE O. KRUEGER
Stanford University

MARTIN BRONFENBRENNER

DISTINGUISHED FELLOW

1997

Martin Bronfenbrenner once said about himself, "'Doubtless I shall end, if I live long enough, 'knowing nothing about everything,' as against the specialist's 'knowing everything about nothing.'" Indeed, Professor Bronfenbrenner's interests spanned several important areas of economics and even beyond it. His contributions are in areas ranging from income distribution theory, monetary economics, fiscal economics, business-cycle theory, imperfect competition, labor economics, and production theory, on one hand, to economic development, Marxian economics, comparative economic systems, history of economic thought, and international economics, on the other. His interest in the development of the Japanese economy deserves special mention. In today's world of specialization, Professor Bronfenbrenner towered over the discipline as a member of that fast-disappearing class of generalists.

His most influential papers include "Some Fundamentals of Liquidity Theory," *Quarterly Journal of Economics* (May 1945); "Price Control Under Imperfect Competition," *American Economic Review* (March 1947); "A Contribution to the Aggregative Theory of Wages," *Journal of Political Economy* (October 1957); "Notes on the Elasticity of Derived Demand," *Oxford Economic Papers* (October 1961); and his landmark survey with Franklyn D. Holzman, "Survey of Inflation Theory," *American Economic Review* (September 1963). Among his several books, *Income Distribution Theory* (Chicago: Aldine Atherton, 1971), continues to be an authoritative work in its area. His macroeconomics textbook, *Macroeconomic Alternatives* (Arlington Heights, IL: AHM Publishing Corporation, 1979), remains unparalleled in its treatment of alternative economic systems.

Professor Bronfenbrenner was born in 1914 in the United States of immigrant parents. He received an A.B. degree from Washington University (St. Louis) and a Ph.D. in Economics from the University of Chicago. The Second World War saw him learning Japanese to translate captured Japanese documents and interrogate Japanese POWs. This ultimately began his long, exciting, and extremely fruitful association with Japan. For almost half a century, Professor Bronfenbrenner held teaching and research positions in several universities in the United States and Japan. In his own words, "the years . . . have been spent on campuses and in airplanes between campuses. . . ." He was a Professor of Economics at Carnegie Mellon University between 1962 and 1971, held the Kenan Chair in Economics at Duke University between 1971 and 1984, and was Professor of International Economics at the Aoyuma Gakuin University at Tokyo, Japan since 1984.

A short time prior to his death in 1997, Martin Bronfenbrenner had been informed that he was elected a Fellow of the American Economic Association.

Martin Bronfenbrenner

INDIAN INSTITUTE OF MANAGEMENT CALCUTTA

HE AMERICAN ECONOMIC REVIEW

or
EY ASHENFELTER

Editors
NIS N. EPPLE
RESTON McAFEE
TTHEW D. SHAPIRO

aging Editor
NN FLEISHER

rd of Editors
VID P. BARON
ANTO BASU
EODORE C. BERGSTROM
OTHY J. BESLEY
NCINE D. BLAU
RLES C. BROWN
AN DRAZEN
OTHY S. FUERST
FREY C. FUHRER
N FULLERTON
DI GALÍ
RTIN S. GAYNOR
RDON H. HANSON
MARK ISAAC
AM B. JAFFE
UL L. JOSKOW
ACY R. LEWIS
RDRE McCLOSKEY
UL R. MILGROM
BERT A. MOFFITT
RISTINA H. PAXSON
LFGANG PESENDORFER
LERIE A. RAMEY
GIO T. REBELO
NIFER F. REINGANUM
ER C. REISS
HARD E. ROMANO
VID H. ROMER
VID E. M. SAPPINGTON
DREW R. SCHOTTER
Y SOLON
BERT W. STAIGER
TIS R. TAYLOR
KIP VISCUSI
RL E. WALSH
VID N. WEIL
NETH D. WEST
HAEL D. WOODFORD
DALL WRIGHT

June 1998

VOLUME 88, NUMBER 3

Articles

Applications of Option-Pricing Theory: Twenty-Five Years Later *Robert C. Merton* 323

Derivatives in a Dynamic Environment *Myron S. Scholes* 350

Unraveling in Matching Markets *Hao Li and Sherwin Rosen* 371

Motivation and Markets *W. Bentley MacLeod and James M. Malcomson* 388

The Economics of Child Labor *Kaushik Basu and Pham Hoang Van* 412

A Theory of Holdouts in Wage Bargaining *Wulong Gu and Peter Kuhn* 428

Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size *Joshua D. Angrist and William N. Evans* 450

Does European Unemployment Prop Up American Wages? National Labor Markets and Global Trade *Donald R. Davis* 478

Growth Cycles *George W. Evans, Seppo Honkapohja, and Paul Romer* 495

Financial Intermediation and Regime Switching in Business Cycles *Costas Azariadis and Bruce Smith* 516

Stock Markets, Banks, and Economic Growth *Ross Levine and Sara Zervos* 537

Financial Dependence and Growth *Raghuram G. Rajan and Luigi Zingales* 559

Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks *Stephen Morris and Hyun Song Shin* 587

Shorter Papers

Compensating Differentials for Gender-Specific Job Injury Risks *Joni Hersch*
Did Unilateral Divorce Raise Divorce Rates? Evidence from Panel Data *Leora Friedberg*
Divorce-Law Changes, Household Bargaining, and Married Women's Labor Supply *Jeffrey S. Gray*
Caps on Political Lobbying *Yeon-Koo Che and Ian L. Gale*
Is There a Free-Rider Problem in Lobbying? Endogenous Tariffs, Trigger Strategies, and the Number of Firms *Paul Pecorino*

Auditors' Report/Audited Financial Statements

- Submit manuscripts (4 copies), single-sided, double-spaced, to:
  Orley Ashenfelter, Editor, *AER*, 209 Nassau Street,
  Princeton, NJ 08542-4607.
- Authorship should be identified only on a removable cover page; the anonymous text should begin on the following page.
- Submission fee: $50 for members; $100 for nonmembers. Please pay with a check or money order payable in United States dollars. Foreign (including Canadian) payments must be in the form of a check drawn on a United States bank payable in United States dollars.
- Style guides will be provided upon request.

It is the policy of the *American Economic Review* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Details of the computations sufficient to permit replication must be provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary, or if, for some other reason, the above requirements cannot be met.

plications of Option-Pricing Theory: Twenty-Five Years Later†

By ROBERT C. MERTON*

he news from Stockholm that the prize in omic sciences had been given for option- ing theory provided unique and signal rec- tion to the rapidly advancing, but still ively new, discipline within economics h relates mathematical finance theory and ce practice.[1] The special sphere of finance in economics is the study of allocation and oyment of economic resources, both spa- y and across time, in an uncertain environ- t. To capture the influence and interaction me and uncertainty effectively requires so- ticated mathematical and computational s. Indeed, mathematical models of modern ce contain some truly elegant applications robability and optimization theory. These lications challenge the most powerful com- tional technologies. But, of course, all that egant and challenging in science need not be practical; and surely, not all that is tical in science is elegant and challenging. we have both. In the time since publi- on of our early work on the option-pricing el, the mathematically complex models of ce theory have had a direct and wide- ing influence on finance practice. This oining of intrinsic intellectual interest with extrinsic application is central to research in modern finance.

It was not always thus. The origins of much of the mathematics in modern finance can be traced to Louis Bachelier's 1900 dissertation on the theory of speculation, framed as an option-pricing problem. This work marks the twin births of both the continuous-time mathematics of stochastic processes and the continuous-time economics of derivative-security pricing. Kiyoshi Itô (1987) was greatly influenced by Bachelier's work in his development in the 1940's and early 1950's of the stochastic calculus, later to become an essential mathematical tool in finance. Paul A. Samuelson's theory of rational warrant pricing, published in 1965, was also motivated by the same piece. However, Bachelier's important work was largely lost to financial economists for more than a half century. During most of that period, mathematically complex models with a strong influence on practice were not at all the hallmarks of finance theory. Before the pioneering work of Markowitz, Modigliani, Miller, Sharpe, Lintner, Fama, and Samuelson in the late 1950's and 1960's, finance theory was little more than a collection of anecdotes, rules of thumb, and shuffling of accounting data. It was not until the end of the 1960's and early 1970's that models of finance in academe become considerably more sophisticated, involving both the intertemporal and uncertainty dimensions of valuation and optimal decision-making. The new models of dynamic portfolio theory, intertemporal capital asset pricing, and derivative-security pricing employed stochastic differential and integral equations, stochastic dynamic programming, and partial differential equations. These mathematical tools were a quantum level more complex than had been used in finance before and they are still the core tools employed today.

The most influential development in terms of impact on finance practice was the Black-Scholes model for option pricing. Yet

---

This article is the lecture Robert C. Merton delivered ockholm, Sweden, December 9. 1997, when he re- d the Alfred Nobel Memorial Prize in Economic Sci- s. The article is copyright © The Nobel Foundation and is published here with the permission of the No- oundation.

Graduate School of Business Administration, Har- University, Boston, MA 02163, and Long-Term Cap- Management, L.P., Greenwich, CT 06831. I am ful to Robert K. Merton, Lisa Meulbroek, and Myron les for their helpful suggestions on this lecture and much more. Over the past 30 years, I have come to an incalculable debt to Paul A. Samuelson, my er, mentor, colleague, co-researcher, and friend. Try have (cf., Merton, 1983, 1992), I cannot find the to pay sufficient tribute to him. I dedicate this lec- Paul and to the memory of Fischer Black.

his section draws on Merton (1994, 1995, 1997b).

paradoxically, the mathematical model was developed entirely in theory, with essentially no reference to empirical option-pricing data as motivation for its formulation. Publication of the model brought the field to almost immediate closure on the fundamentals of option-pricing theory. At the same time, it provided a launching pad for refinements of the theory, extensions to derivative-security pricing in general, and a wide range of other applications, some completely outside the realm of finance. The Chicago Board Options Exchange (CBOE), the first public options exchange, began trading in April 1973, and by 1975, traders on the CBOE were using the model to both price and hedge their option positions. It was so widely used that, in those pre-personal-computer days, Texas Instruments sold a handheld calculator specially programmed to produce Black-Scholes option prices and hedge ratios. That rapid adoption was all the more impressive, as the mathematics used in the model were not part of the standard mathematical training of either academic economists or practitioner traders.

Academic finance research of the 1960's, including capital asset pricing, performance, and risk measurement, and the creation of the first large-scale databases for security prices essential for serious empirical work, have certainly influenced subsequent finance practice. Still the speed of adoption and the intensity of that influence was not comparable to the influence of the option model. There are surely several possible explanations for the different rates of adoption in the 1960's and the 1970's. My hypothesis is that manifest ''need'' determined that difference. In the 1960's, especially in the United States, financial markets exhibited unusually low volatility: the stock market rose steadily, interest rates were relatively stable, and exchange rates were fixed. Such a market environment provided investors and financial service firms with little incentive to adopt new financial technology, especially technology designed to help manage risk. However, the 1970's experienced several events that caused both structural changes and large increases in volatility. Among the more important events were: the shift from fixed to floating exchange rates with the fall of Bretton Woods and the devaluation of the dollar; the world oil-price shock with the creation OPEC; double-digit inflation and interest ra in the United States; and the extraordin real-return decline in the U.S. stock mar from a peak of around 1050 on the Dow Jo Industrial Average in the beginning of 197 about 580 at the end of 1974. As a result, increased demand for managing risks in a v atile and structurally different economic e vironment contributed to the major success the derivative-security exchanges created the 1970's to trade listed options on stoc futures on major currencies, and futures fixed-income instruments. This success in increased the speed of adoption for quant tive financial models to help value options assess risk exposures.

The influence of option-pricing theory finance practice has not been limited to fin cial options traded in markets or even to rivative securities generally. As we shall the underlying conceptual framework ori nally used to derive the option-pricing form can be used to price and evaluate the risk wide array of applications, both financial nonfinancial. Option-pricing technology played a fundamental role in supporting creation of new financial products and mark around the globe. In the present and in the pending future, that role will continue expa ing to support the design of entirely n financial institutions, decision-making by s ior management, and the formulation of pub policy on the financial system. To undersc that point, I begin with a few remarks ab financial innovation of the past, this adumb tion to be followed in later sections with a tailed listing of applications of the opti technology that include some observations the directions of future changes in finan services.

New financial product and market desig improved computer and telecommunicati technology, and advances in the theory of nance during the past quarter century have to dramatic and rapid changes in the struct of global financial markets and institutio The scientific breakthroughs in financial m eling in this period both shaped and w shaped by the extraordinary flow of finan innovation which coincided with th changes. Thus, the publication of the opti cing model in 1973 surely helped the development and growth of the listed options and over-the-counter (OTC) derivatives markets. At the same time, the extraordinary growth and success of these markets just as surely stimulated further development and research focus on the derivative-security pricing models. To see this in perspective, consider some of the innovative changes in market structure and scale of the global financial system since 1973. There occurred the aforementioned fall of Bretton Woods leading to floating exchange rates for currencies; the development of the national mortgage market in the United States which in turn restructured that entire industry; passage of the Employee Retirement Income Security Act (ERISA) in 1974 with the subsequent development of the U.S. pension-fund industry; the first money-market fund with check writing that also took place in 1974; and the explosive growth in mutual fund assets from $48 billion 25 years ago to $4.3 trillion today (a 100-fold increase), with one institution, Fidelity Investments, accounting for some $500 billion by itself. In this same period, average daily trading volume on the New York Stock Exchange grew from 12 million shares to more than 300 million. Even more dramatic were the changes in Europe and in Asia. The cumulative impact has significantly affected all of us—users, producers, or overseers of the financial system.

Nowhere has this been more the case than the development, refinement, and broad-based implementation of contracting technology. Derivative securities such as futures, options, swaps, and other contractual agreements—the underlying substantive instruments for which our model was developed—provide a prime example. Innovations in financial-contracting technology have improved efficiency by expanding opportunities for risk sharing, lowering transactions costs, and reducing information and agency costs. The numbers reported for the global use of derivative securities are staggering (the figure $70 trillion appeared more than once in the news stories surrounding the award of the Nobel Prize and there are a number of world banking institutions with reported multi-trillion-dollar, off-balance-sheet derivative positions). However, since these are notional amounts (and often involve double-counting), they are meaningless for assessing either the importance or the risk exposure to derivative securities.[2] Nevertheless, it is enough to say here that, properly measured, derivatives are ubiquitous throughout the world financial system and that they are used widely by nonfinancial firms and sovereigns as well as by institutions in virtually every part of their financing and risk-managing activities. Some observers see the extraordinary growth in the use of derivatives as fad-like, but a more likely explanation is the vast saving in transactions costs derived from their use. The cost of implementing financial strategies for institutions using derivatives can be one-tenth to one-twentieth of the cost of executing them in the underlying cash-market securities.[3] The significance of reducing spread costs in financing can be quite dramatic for corporations and for sovereigns: for instance, not long ago, a 1-percent (i.e., 100-basis-point) reduction in debt-spread cost on Italian government debt would have reduced the deficit by an amount equal to 1.25 percent of the gross domestic product of Italy.

Further improved technology, together with growing breadth and experience in the applications of derivatives, should continue to reduce transactions costs as both users and producers of derivatives move along the learning curve. Like retail depositors with automatic-teller machines in banks, initial resistance by institutional clients to contractual agreements can be high, but once customers use them they tend not to return to the traditional alternatives for implementing financial strategies.

A central process in the past two decades has been the remarkable rate of globalization of the financial system. Even today, inspection of the diverse financial systems of individual nation-states would lead one to question how effective integration across geopolitical borders could have realistically taken place since those systems are rarely compatible in institutional forms, regulations, laws, tax structures, and business practices. Still, significant integration did take place. This was made possible in large part by derivative securities functioning as "adapters." In general, the flexibility created by the widespread use of contractual agreements, other derivatives, and specialized institutional designs provides an offset to dysfunctional institutional rigidities.[4] More specifically, derivative-security contracting technologies provide efficient means for creating cross-border interfaces among otherwise incompatible domestic systems, without requiring widespread or radical changes within each system. For that reason, implementation of derivative-security technology and markets within smaller and emerging-market countries may help form important gateways of access to world capital markets and global risk sharing. Such developments and changes are not limited only to the emerging-market countries with their new financial systems. Derivatives and other contracting technologies are likely to play a significant role in the financial engineering of the major transitions required for European Monetary Union and for the major restructuring of financial institutions in Japan.

---

[2] Notional amounts typically represent either the total value of the underlying asset on which payments on the derivative is determined (e.g., interest rate swap contracts) or the exercise price on an option. The value of the derivative contract itself is often a small fraction of its notional amount.

[3] See André F. Perold (1992) for a case study illustrating the savings in transactions costs, taxes, and custodial fees from using derivatives instead of the cash market. Myron S. Scholes (1976) provides an early analysis of the effect of taxes on option prices.

With this introduction as background, I turn now to the key conceptual and mathematical framework underlying the option-pricing model and its subsequent applications.

I. General Derivation of Derivative-Security Pricing

I understand that it is customary in these lectures for the Nobel Laureates to review the background and the process leading up to their discoveries. Happily, there is no need to do so here since that has been done elsewhere in Fischer Black (1989), Peter L. Bernstein (1992 Ch. 11), Merton and Scholes (1995), and Scholes (1998). Instead, I briefly summarize. My principal contribution to the Black-Scholes option-pricing theory was to show that the dynamic trading strategy prescribed by Black and Scholes to offset the risk exposure of an option would provide a perfect hedge in the limit of continuous trading. That is, if one could trade continuously without cost, then following their dynamic trading strategy using the underlying traded asset and the riskless asset would exactly replicate the payoffs on the option. Thus, in a continuous-trading financial environment, the option price must satisfy the Black-Scholes formula or else there would be an opportunity for arbitrage profits. To demonstrate this limit-case result, I applied the tools developed in my earlier work (1969, 1971) on the continuous-time theory of portfolio selection. My 1973a paper also extended the applicability of the Black-Scholes model to allow for stochastic interest rates on the riskless asset, dividend payments on the underlying asset, a changing exercise price, American-type early-exercise of the option, and other "exotic" features such as the "down-and-out" provision on the option. I am also responsible for naming the model, "the Black-Scholes Option-Pricing Model."[5]

The derivations of the pricing formula in both of our 1973 papers make the following assumptions:

---

[4] Scholes and Mark A. Wolfson (1992) develop the principles of security and institutional design along these lines. See also Perold (1992) and Merton (1993, 1995). Inspection of the weekly *International Financing Review* will find the widespread and varied applications of financial engineering, derivatives, special-purpose vehicles, and securities for private-sector and sovereign financing in every part of the world.

[5] My 1970 working paper was the first to use the "Black-Scholes" label for their model (cf., Merton 1992 p. 379). This same paper was given at the July 1970 Wells Fargo Capital Market Conference, since made "famous" (or notorious) by Bernstein (1992 p. 223) as the one at which I ". . . inconveniently overslept . . ." the morning session and missed the Black and Scholes presentation. The second instance naming their model was in the 1971 working paper version of Merton (1973a). Samuelson (1972) is the first published usage: both in the main text and in my Appendix to that paper which derives the model and refers to it as the "Black-Scholes formula." The formula is cited in Roger J. Leonard (1971) and Carliss Baldwin (1972), the earliest theses to apply the model. Somewhat ironically, all these references to the "Black-Scholes model" appear before the actual publication of either Black and Scholes (1972) or (1973).

ASSUMPTION 1: *"Frictionless" and "continuous" markets—There are no transactions costs or differential taxes. Markets are open all the time and trading takes place continuously. Borrowing and short-selling are allowed without restriction. The borrowing and lending rates are equal.*

ASSUMPTION 2: *Underlying asset-price dynamics—Let $V = V(t)$ denote the price at time t of a limited-liability asset, such as a share of stock. The posited dynamics for the instantaneous returns can be described by an Itô-type stochastic differential equation with continuous sample paths given by*

$$dV = [\alpha V - D_1(V, t)]\,dt + \sigma V dZ,$$

where: $\alpha \equiv$ instantaneous expected rate of return on the security; $\sigma^2 \equiv$ instantaneous variance rate, which is assumed to depend, at most, on $V(t)$ and t (i.e., $\sigma^2 = \sigma^2(V, t)$); dZ is a Wiener process; and $D_1 \equiv$ dividend payment flow rate. With limited liability, to avoid arbitrage, $V(t) = 0$ for all $t \geq t^*$ if $V(t^*) = 0$. Hence D_1 must satisfy $D_1(0, t) = 0$. Other than a technical requirement of bounded variation, α can follow a quite general stochastic process, dependent on V, other security prices, or state variables. In particular, the assumed dynamics permit a mean-reverting process for the underlying asset's returns.

ASSUMPTION 3: *Default-free bond-price dynamics—Bond returns are assumed to be described by Itô stochastic processes with continuous sample paths. In the original Black and Scholes formulation and for exposition convenience here, it is assumed that the riskless instantaneous interest rate, $r(t) = r$, is a constant over time.*

ASSUMPTION 4: *Investor preferences and expectations—Investor preferences are assumed to prefer more to less. All investors are assumed to agree on the function σ^2 and on the Itô process characterization for the return dynamics. It is not assumed that they agree on the expected rate of return, α.*

ASSUMPTION 5: *Functional dependence of the option-pricing formula—The option price is assumed to be a twice-continuously differentiable function of the asset price, V, default-free bond prices, and time.*

In the particular case of a nondividend-paying asset ($D_1 = 0$) and a constant variance rate, σ^2, these assumptions lead to the Black-Scholes option-pricing formula for a European-type call option with exercise price L and expiration date T, written as

$$(1)\quad C(V,t) = VN(d) - L\exp(-r[T-t]) \times N(d - \sigma\sqrt{T-t}),$$

where $d = (\ln[V/L] + [r + \sigma^2/2][T - t])/\sigma\sqrt{T-t}$ *and $N(\;)$ is the cumulative density function for the standard normal distribution.*

Subsequent research in the field proceeded along three dimensions: applications of the technology to other than financial options (which is discussed in the next section); empirical testing of the pricing formula, which began with a study using over-the-counter data from a dealer's book obtained by Black and Scholes (1972); and attempts to weaken the assumptions used in the derivation, and thereby to strengthen the foundation of the applications developed from this research. The balance of this section addresses issues of the latter dimension.

Early concerns raised about the model's theoretical foundation came from John B. Long (1974) and Clifford W. Smith, Jr. (1976), who questioned Assumption 5: namely, how does one know that the option prices do not depend on other variables than the ones assumed (for instance, the price of beer), and why should the pricing function be twice-continuously differentiable? These concerns were resolved in an alternative derivation in Merton (1977b) which shows that Assumption 5 is a derived consequence, not an assumption, of the analysis.[6]

---

[6] As another instance of early questioning of the core model, a paper I refereed argued that Black-Scholes must be fundamentally flawed because a different valuation formula is derived from the replication argument if the R. L. Stratonovich (1968) stochastic calculus is used for mod-

A broader, and still open, research issue is the robustness of the pricing formula in the absence of a dynamic portfolio strategy that exactly replicates the payoffs to the option security. Obviously, the conclusion on that issue depends on why perfect replication is not feasible as well as on the magnitude of the imperfection. Continuous trading is, of course, only an idealized prospect, not literally obtainable; therefore, with discrete trading intervals, replication is at best only approximate. Subsequent simulation work has shown that within the actual trading intervals available and the volatility levels of speculative prices, the error in replication is manageable, provided, however, that the other assumptions about the underlying process obtain. John C. Cox and Stephen A. Ross (1976) and Merton (1976a, b) relax the continuous sample-path assumption and analyze option pricing using a mixture of jump and diffusion processes to capture the prospect of nonlocal movements in the underlying asset's return process.[7] Without a continuous sample path, replication is not possible and that rules out a strict no-arbitrage derivation. Instead, the derivation of the option-pricing model is completed by using equilibrium asset pricing models such as the Intertemporal CAPM (Merton, 1973b) and the Arbitrage Pricing Theory (Ross, 1976a).[8] This approach relates back to the original way in which Black and Scholes derived their model using the classic Sharpe-Lintner CAPM.[9] There has developed a considerable literature on the case of imperfect replication [cf., Breeden (1984), Hans Föllmer and Dieter Sondermann (1986), Steven Figlewski (1989), Dimitris Bertsimas et al. (1997), Mark H. A. Davis (1997), and Marc Romano and Nizar Touzi (1997)].

On this occasion, I reexamine the imperfect-replication problem for a derivative security linked to an underlying asset that is not continuously available for trading in an environment in which some assets are tradable at any time. As is discussed in the section to follow, nontradability is the circumstance for several important classes of applications that have evolved over the last quarter century, which include among others, the pricing of financial guarantees such as deposit and pension insurance and the valuation of nonfinancial or "real" options. Since the Black-Scholes model was derived by assuming that the underlying asset is continuously traded, questions have been raised about whether the pricing formula can be properly applied in those applications. The derivation follows along the lines presented in Merton (1977b, 1997b) for the perfect-replication case.

A *derivative security* has contractually determined payouts that can be described by functions of observable asset prices and time. These payout functions define the derivative. The terms are expressed as follows:

Let $W(t)$

= price of a derivative security at time t.

$$(2)\quad \text{If } V(t) \geqq \bar{V}(t) \text{ for } 0 \leq t < T, \quad \text{then } W(t) = f[V(t), t];$$

$$\text{If } V(t) \leqq \underline{V}(t) \text{ for } 0 \leq t < T, \quad \text{then } W(t) = g[V(t), t];$$

$$\text{If } t = T, \quad \text{then } W(T) = h[V(T)].$$

---

eling instead of the Itô calculus. My report showed that while the paper's mathematics was correct, its economics was not: A Stratonovich-type formulation of the underlying price process implies that traders have a partial knowledge about future asset prices that the nonanticipating character of the Itô process does not. The "paradox" is thus resolved because the assumed information sets are essentially different and hence, so should the pricing formulas.

[7] Since a discontinuous sample-path price process for the underlying asset rules out perfect hedging even with continuous trading but a continuous-sample-path process with stochastic volatility does not, there is considerable interest in testing which process fits the data better. See Eric Rosenfeld (1980), an early developer of such tests, and James B. Wiggins (1987).

[8] The important Douglas T. Breeden (1979) Consumption-based Capital Asset Pricing Model, which was not published at the time of these papers, can also be used to complete those models.

[9] See Black (1989) and Scholes (1998). Fischer Black always maintained with me that the CAPM-version of the option-model derivation was more robust because continuous trading is not feasible and there are transactions costs. As noted in Merton (1973a p. 161), the discrete-time Samuelson-Merton (1969) model also gives the Black-Scholes formula under special conditions.

For $0 \le t \le T$, the derivative security receives a payment flow rate specified by $D_2(V, t)$. The terms as described in (2) are to be interpreted as follows: the first time that $V(t) \geqq \bar{V}(t)$ or $V(t) \leqq \underline{V}(t)$, the owner of the derivative must exchange it for cash according to the schedule in (2). If no such events occur for $t < T$, then the security is redeemed at $t = T$ for cash according to (2). T is called the *maturity date* (or expiration date, or redemption date) of the derivative. The derivative security is thus defined by specifying the contingent payoff functions f, g, h, D_2, and T. In some cases, the schedules or the boundaries $\bar{V}(t)$ and $\underline{V}(t)$ are contractually specified; in others, they are determined endogenously as part of the valuation process, as in the case of the early-exercise boundary for an American-type option.

By arbitrage restrictions, the derivative security will have *limited liability* if and only if $g \ge 0$, $h \ge 0$, $f \ge 0$, and $D_2(0, t) = 0$.

If (as drawn in Figure 1) the boundaries $\underline{V}(t)$, and $\bar{V}(t)$ are continuous functions, then because $V(t)$ has a continuous sample path in t by Assumption 2, one has that: (i) if $V(t) < \underline{V}(t)$ for some t, then there is a $\underline{t}$, $\underline{t} < t$, so that $V(\underline{t}) = \underline{V}(\underline{t})$; and (ii) if $V(t) > \bar{V}(t)$ for some t, then there is a $\bar{t}$, $\bar{t} < t$, so that $V(\bar{t}) = \bar{V}(\bar{t})$. Hence, in this case, the inequalities for V can be neglected in (2) and the only relevant region for analysis is $\underline{V}(t) \le V(t) \le \bar{V}(t)$, $0 \le t \le T$.

With the derivative-security characteristics fully specified, we turn now to the fundamental production technology for hedging the risk of issuing a derivative security and for evaluating the cost of its production. To locate the derivation in a more substantive framework, I posit a hypothetical financial intermediary that creates derivative securities in principal transactions for its customers by selling them contracts which are its obligation. It uses the capital markets or transactions with other institutions to hedge the contractual liabilities so created by dynamically trading in the underlying securities following a strategy designed to reproduce the cash flows of the issued contracts as accurately as it can. If the intermediary cannot perfectly replicate the payoffs to the issued derivative, it either obtains adequate equity to bear the residual risks of its imperfectly hedged positions or it securitizes those positions by bundling them into a portfolio for a special-purpose financial vehicle which it then sells either in the capital market or to a consortium of other institutions in a process similar to the traditional reinsurance market. Although surely a caricature, the following description is nevertheless not far removed from real-world practice.

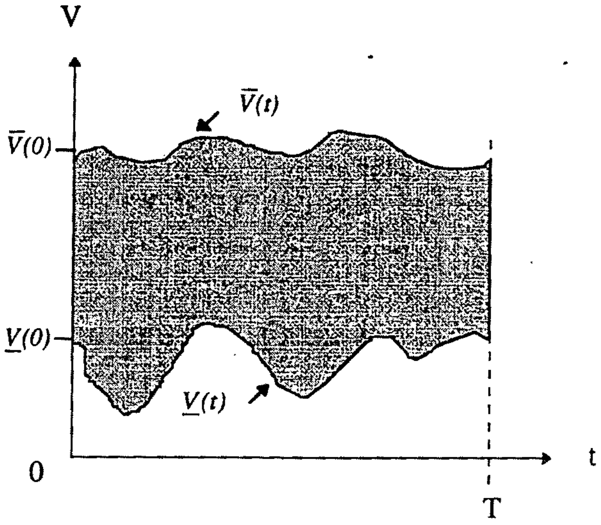


FIGURE 1. RELEVANT REGION OF V: $\underline{V}(t) \le V(t) \le \bar{V}(t)$, $0 \le t \le T$

The objective is to find a feasible, continuous-trading portfolio strategy constructed from all available traded assets including the riskless asset that comes "closest" to satisfying the following four properties: if $P(t)$ denotes the value of the portfolio at time t, then for $0 \le t \le T$:

(i) at t, if $V(t) = \underline{V}(t)$, then $P(t) = g[\underline{V}(t), t]$;
(ii) at t, if $V(t) = \bar{V}(t)$, then $P(t) = f[\bar{V}(t), t]$;
(iii) for each t, the payout rate on the portfolio is $D_2(V, t)dt$;
(iv) at $t = T$, $P(T) = h[V(T)]$.

Call this portfolio the "hedging portfolio" for the derivative security defined by (2). That portfolio is labelled as "portfolio (*)." In the special, but important, case in which the portfolio meets the above conditions exactly, the hedging portfolio is called the "replicating portfolio" for the derivative security.

Bertsimas et al. (1997) study the complementary problem of "closeness" of dynamic replication where they assume that one can trade in the underlying asset but that trading is not continuous. They apply stochastic dynamic programming to derive optimal strategies to minimize mean-squared tracking error. These strategies are then employed in simulations to estimate quantitatively how close one can get to dynamic completeness.

I determine the optimal hedging portfolio in two steps: first, find the portfolio strategy constructed from all continuously traded assets that has the smallest "tracking error" in replicating the returns on the underlying asset. For the underlying asset with price V, I call this portfolio, the "V-Fund." In the second step, derive the hedging portfolio for the derivative security as a dynamic portfolio strategy mixing the V-Fund with the riskless asset.

Let $S_i(t)$ denote the price of continuously traded asset i at time t. There are n such risky assets plus the riskless asset which are traded continuously. The dynamics for S_i are assumed to follow a continuous-sample-path Itô process given by

$$(3)\qquad dS_i = \alpha_i S_i\,dt + \sigma_i S_i\,dZ_i, \quad i = 1, \dots, n,$$

where α_i is the instantaneous expected rate of return on asset i; dZ_i is a Wiener process; σ_{ij} is the instantaneous covariance between the returns on i and j [that is, $(dS_i/S_i)(dS_j/S_j) = \sigma_{ij}dt$ and $\sigma_{ii} = \sigma_i^2$]; let η_i be defined as the instantaneous correlation between dZ_i and dZ in Assumption 2 such that $dZ_i\,dZ = \eta_i\,dt$. Let $S(t)$ denote the value of the V-Fund portfolio and let $w_i(t)$ denote the fraction of that portfolio allocated to asset i, $i = 1, \dots, n$, at time t. The balance of the portfolio's assets are invested in the riskless asset. The dynamics for S can be written as

$$(4)\qquad dS = [\mu S - D_1(V, t)]\,dt + \delta S\,dq,$$

where $\mu = r + \sum_{i=1}^{n} w_i(t)[\alpha_i - r]$, $\delta^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i(t)w_j(t)\sigma_{ij}$, and $dq = [\sum_{i=1}^{n} w_i(t)\sigma_i\,dZ_i]/\delta$.

To create the V-Fund, the w_i are chosen so as to minimize the unanticipated part of the difference between the return on the underlying asset and the traded portfolio's return. That is, at each point in time, the portfolio allocation is chosen so as to minimize the instantaneous variance of $[dS/S - dV/V]$. As shown in Merton (1992, Theorem 15.3 p. 501), the portfolio rule that does this is given by

$$(5)\qquad w_i(t) = \sigma \sum_{k=1}^{n} v_{ki}\sigma_k\eta_k, \qquad i = 1, \dots, n,$$

where v_{ki} is the kth-ith element of the inverse of the variance-covariance matrix of the returns on the n risky continuously traded assets. From Merton (1992 p. 502), the instantaneous correlation between the returns on the V-Fund and the underlying asset, $\rho\,dt = dZ\,dq$, can be written as

$$(6)\qquad \rho = \left(\sum_{k=1}^{n}\sum_{i=1}^{n} v_{ki}\sigma_k\sigma_i\eta_k\eta_i\right)^{1/2}$$

and

$$(7)\qquad \delta = \rho\sigma.$$

The dynamics of the tracking error can thus be written as

$$(8)\qquad dS/S - dV/V = (\mu - \alpha)\,dt + \theta\,db,$$

where $\theta^2 = (1 - \rho^2)\sigma^2$ and the Wiener process $db = (\rho\,dq - dZ)/\sqrt{1 - \rho^2}$. As shown in Merton (1992 equation 15.51), it follows that

$$(9)\qquad dS_i/S_i\,db = 0, \qquad i = 1, \dots, n.$$

That is, the tracking error in (8) is uncorrelated with the returns on all traded assets, which is a consequence of picking the portfolio strategy that minimizes that error.

With this, we now proceed with a "cookbook-like" derivation of the production process for our hypothetical financial intermediary to best hedge the cash flows of the derivative securities it issues. The derivation begins with a description of the activities for the intermediary's quantitative-analysis ("quant") department which is responsible for gathering the variance-covariance information necessary to use (5) to construct and ntain the V-Fund portfolio. It is also as-ned the responsibility to solve the following ar parabolic partial differential equation $F[V, t]$:

$$0 = \tfrac{1}{2}\sigma^2(V,t)V^2F_{11}[V,t] + [rV - D_1(V,t)]F_1[V,t] - rF[V,t] + F_2[V,t] + D_2(V,t)$$

ject to the boundary conditions: for $\underline{V}(t) \leq V \leq \bar{V}(t)$ and $t < T$,

$$F[\bar{V}(t), t] = f[\bar{V}(t), t] \geq 0; \tag{11}$$

$$F[\underline{V}(t), t] = g[\underline{V}(t), t] \geq 0; \tag{12}$$

$$F[V, T] = h[V] \geq 0, \tag{13}$$

here $F_{11} \equiv \partial^2F/\partial V^2$, $F_1 \equiv \partial F/\partial V$; and $\equiv \partial F/\partial t$. Note that the nonnegativity con-tions in (11)–(13) together with $D_2(0, t) =$ implies that the derivative security has lim-d liability. As a mathematical question, this a well-posed problem, and a solution to 0)–(13) exists and is unique.

Having solved for the function $F[V, t]$, the ant department has the prescribed ongoing ks at each time t $(0 \leq t \leq T)$ to:

(i) ask the trading desk for the prices of all traded assets necessary to determine the price $S(t)$ of the V-Fund and the best estimate of the current price of the underlying asset, $V(t)$;

(ii) compute from the solution to (10)–(13)

$$M(t) \equiv F_1[V(t), t]V(t);$$

(iii) tell the trading desk that the strategy of portfolio (*) requires that $\$M(t)$ be invested in the V-Fund for the period t to $t + dt$;

(iv) compute $Y(t) \equiv F[V(t), t]$ and store $Y(t)$ in the intermediary's data files for (later) analysis of the time series (i.e., stochastic process) $Y(t)$.

The prescription for the execution or ading-desk activities of the intermediary is as follows: At time $t = 0$, give the trading desk $\$P(0)$ as an initial funding (investment) for portfolio (*) which contains the V-Fund asset and the riskless asset. Let $P(t)$ denote the value of portfolio (*) at t, after having made any prescribed cash distribution (payment) from the portfolio. The trading desk has the job at each time t $(0 \leq t \leq T)$ to:

(a) determine the current prices of the underlying asset, $V(t)$ and all individual traded assets held in the V-Fund, and send that price information to the quant department;

(b) pay a cash distribution of $\$D_2[V(t), t]dt$ to the customer holding the derivative security, by selling securities in the portfolio (if necessary);

(c) compute the value of the balance of the portfolio, $P(t)$;

(d) receive instructions on $M(t)$ from the quant department;

(e) readjust the portfolio allocation so that $\$M(t)$ is now invested in the V-Fund and $\$[P(t) - M(t)]$ is invested in the riskless asset.

It follows that the dynamics for the value of portfolio (*) are given by

$$dP = M(t)\frac{dS}{S} + M(t)\frac{D_1(V,t)}{S}dt + [P - M(t)]rdt - D_2(V,t)dt \tag{14}$$

where

$$M(t)\frac{dS}{S} = \text{price appreciation;}$$

$$M(t)\frac{D_1(V,t)}{S}dt = \text{dividend payments received into the portfolio;}$$

$$[P - M(t)]rdt = \text{interest earned by the portfolio;}$$

$$D_2(V,t)dt = \text{cash distribution to customer.}$$

Noting that $M(t) = F_1[V, t]V$, one has by substitution from (4) into (14) that the dynamics of P satisfy

$$\begin{aligned} dP &= F_1[V, t]V dS/S \\ &\quad + F_1[V, t]V D_1(V, t)/S \\ &\quad + (P - F_1[V, t]V)r dt \\ &\quad - D_2(V) dt \\ &= [F_1 V(\mu - r) + rP - D_2]\, dt \\ &\quad + F_1 V \delta dq. \end{aligned} \tag{15}$$

Return now to the quant department to derive the dynamics for $Y(t)$. From (iv), one has that $Y(t) = F[V, t]$ for $V(t) = V$. Because F is the solution to (10)–(13), F is a twice-continuously differentiable function of V and t. Therefore, we can apply Itô's lemma, so that for $V(t) = V$,

$$\begin{aligned} dY &= F_1[V, t]dV + F_2[V, t]dt \\ &\quad + \tfrac{1}{2}F_{11}[V, t](dV)^2 \\ &= [\tfrac{1}{2}\sigma^2 V^2 F_{11} + F_1(\alpha V - D_1) \\ &\quad + F_2]\, dt + F_1 V \sigma dZ \end{aligned} \tag{16}$$

because $(dV)^2 = \sigma^2 V^2 dt$. Because $F[V, t]$ satisfies (10), one has that

$$\begin{aligned} &\tfrac{1}{2}\sigma^2 V^2 F_{11} - D_1 F_1 + F_2 \\ &\quad = rF - rVF_1 - D_2. \end{aligned} \tag{17}$$

Substituting (17) into (16), one can rewrite (16) as

$$\begin{aligned} dY &= [F_1(\alpha - r)V + F - D_2]\, dt \\ &\quad + F_1 V \sigma dZ. \end{aligned} \tag{18}$$

Note that the calculation of $Y(t)$ and its dynamics by the quant department in no way requires knowledge of the time series of values for portfolio (*), $\{P(t)\}$, that are calculated by the trading desk. Putting these two time series together, we define $Q(t) \equiv P(t) - Y(t)$. It follows that $dQ = dP - dY$. Substituting for dP from (15) and for dY from (18), rearranging terms using (8), one has that

$$\begin{aligned} dQ &= rQdt + F_1 V(dS/S - dV/V) \\ &= (rQ + F_1 V[\mu - \alpha])\, dt \\ &\quad + F_1 V \theta db. \end{aligned} \tag{19}$$

At this point, I digress to examine the special case in which perfect replication of the return on the underlying asset obtains (i.e., $\rho = 1$ and there is no tracking error). In that case, equation (19) reduces to an ordinary differential equation ($\dot{Q}/Q = r$) with solution

$$Q(t) = Q(0)\exp(rt) \tag{20}$$

where $Q(0) = P(0) - Y(0) = P(0) - F[V(0), 0]$. Therefore, if the initial funding provided to the trading desk for portfolio (*) is chosen so that $P(0) = F[V(0), 0]$, then from (20), $Q(t) \equiv 0$ for all t and

$$P(t) = F[V(t), t]. \tag{21}$$

By comparison of (11)–(13) with (2), one has from (21) that the (*)-portfolio strategy generates the identical payment flows and terminal (and boundary) values as the derivative security described at the outset of this analysis. That is, for a one-time, initial investment of $\$F[V(0), 0]$, a feasible portfolio strategy has been found that exactly replicates the payoffs to the derivative security. Thus, $\$F[V(0), 0]$ is the cost to the intermediary for producing the derivative. *If the derivative security is traded,* then to avoid ("conditional") arbitrage (conditional on σ, r, D_1), its price must satisfy

$$W(t) = P(t) = F[V(t), t]. \tag{22}$$

Since the absence of arbitrage opportunities is a *necessary* condition for equilibrium, it follows that equilibrium prices for derivative securities on continuously tradable underlying assets must satisfy (22). This is, of course, the original Black-Scholes result and the V-Fund degenerates into a single asset, the underlying asset itself. However, note that (22) obtains without assuming that the derivative-pricing function is a twice-continuously differentiable function of V and t. The smoothness of the pricing function is instead a derived conclusion.

Note further that the development of the (*)-portfolio strategy did not require that the derivative security [defined by (2)] actually trades in the capital market. The (*)-portfolio strategy provides the technology for "manufacturing" or synthetically creating the cash flows and payoffs of the derivative security if it does not exist. That is, if one describes a state-contingent schedule of outcomes for a portfolio [i.e., specifies f, g, h, D_2, T, $\underline{V}(t)$, $\bar{V}(t)$], then the (*)-portfolio strategy provides the trading rules to create this pattern of payouts and it specifies the cost of implementing those rules. The cost of creating the security at time t is thus $F[V(t), t]$. Moreover, if the financial services industry is competitive, then price equals marginal cost, and (22) obtains as the formula for equilibrium prices of derivatives sold directly by intermediaries.

Returning from this digression to the case of imperfect replication, one has, by construction of the process for Y, that $Q = P - Y$ is the cumulative arithmetic tracking error for the hedging portfolio. By inspection of (19), the instantaneous tracking error for the derivative security is perfectly correlated with the tracking error of the V-Fund. Hence, from (9), it follows that the tracking error for the hedging portfolio is uncorrelated with the returns on all continuously traded assets. Using this lack of correlation with any other traded asset, I now argue that in this case the replication-based valuation can be used for pricing the derivative security even though replication is not feasible.

As we know, in all equilibrium asset-pricing models, assets that have only nonsystematic or diversifiable risk are priced to yield an expected return equal to the riskless rate of interest. The condition satisfied by the tracking-error component of the hedging portfolio satisfies an even stronger no-correlation condition than either a zero-beta asset in the CAPM, a zero multibeta asset of the Intertemporal CAPM, or a zero factor-risk asset of the Arbitrage Pricing Theory. Thus, by any of those theories, the equilibrium condition from either (8) or (19) is that

$$\mu = \alpha. \tag{23}$$

If (23) obtains, it follows immediately that the equilibrium price for the derivative security is $F[V(t), t]$, the same formula "as if" the underlying asset is traded continuously. And as a consequence, the Black-Scholes formula would apply even in those applications in which the underlying asset is not traded.

As is well known from the literature on incomplete markets, (23) need not obtain if the creation of the new derivative security helps complete the market for a large enough subset of investors that the incremental dimension of risk spanned by this new instrument is "priced" as a systematic risk factor with an expected return different from the riskless interest rate. Markets tend to remain incomplete with respect to a particular risk either because the cost of creating the securities necessary to span that risk exceeds the benefits, or because nonverifiability, moral-hazard, or adverse-selection problems render the viability of such securities untenable. Generally, major macro risks for which significant pools of investors want to manage their exposures are not controllable by any group of investors, and it is unlikely that any group would have systematic access to materially better information about those risks. Hence, the usual asymmetric-information and incentive reasons given for market failure do not seem to be present. In systems with well-developed financial institutions and markets and with today's financial technology, it is thus not readily apparent what factors make the cost of developing standardized derivative markets (e.g., futures, swaps, options) prohibitive if, in large scale, there is a significant premium latently waiting to be paid by investors who currently participate in the markets. On a more prosaic empirical note, in most applications of the option-pricing model, the "residual" or tracking-error variations are likely to be specific to the underlying project, firm, institution, or person, and thereby they are unlikely candidates for macro-risk surrogates. These observations support the prospects for (23) to obtain.

However, the risk need not be macro in scope in order to be significant to one investor or a small group of investors. Obvious examples of such risks would be various firm- or person-specific components of human capital, including death and disability risks. To make a case for instruments with these types of exposures to be priced with a risk premium, incomplete-market models often focus on the "incipient-demand" (or "maximum reservation") price or risk premium that an investor would pay to eliminate a risk that is not covered in the market by the existing set of securities. In the abstract, that price, of course, can be quite substantial. However, arguments along these lines to explain financial product pricing implicitly assume a rather modest and static financial services sector. A classic example is life insurance. Risk-averse individuals with families may, if necessary, be willing to pay a considerable premium for life insurance, well in excess of the actuarial mortality risk, even after taking into account moral-hazard and person-specific informational asymmetries. Moreover, if the analysis further postulates a financial sector so crude that bilateral contracts between risk-averse individuals are the only way to obtain such insurance, then the equilibrium price for such insurance in that model can be so large that few, if any, contracts are created. But, such models are a poor descriptor of the real world. If the institutions and markets were really that limited, the incentives for change and innovation would be enormous. Modern finance technology and experience in implementing it provide the means for such change. And if, instead, one admits into the model just the classic mechanism for organizing an "insurance" institution (whether government-run or private sector) to take advantage of the enormous diversification benefits of pooling such risks and subdividing them among large numbers of participants, then the equilibrium price equals the "supply" price of such insurance contracts which approaches the actuarial rate.

As is typical in analyses of other industries, the equilibrium prices of financial products and services are more closely linked to the costs of the efficient producers than to the inefficient ones (except perhaps as a very crude upper bound to those prices). Furthermore, the institutional structure of the financial system is neither exogenous nor fixed. In theory and in practice, that structure changes in response to changing technology and to profit opportunities for creating new products and existing products more efficiently. As discussed at length elsewhere (Merton, 1992 pp. 457–67, 535–36), a financial sector with a rich and well-developed structure of institutions can justify a "quasi-dichotomy" modeling approach to the pricing of real and financial assets that employs "reduced-form" equilibrium models with a simple financial sector in which all agents are assumed to be minimum-cost information processors and transactors. However, distortions of insights into the real world can occur if significant costs for the agents are introduced into the model while the simple financial sector is retained as an unchanged assumption. Put simply, high transaction and information costs for most of the economy's agents to directly create their own financial products and services does not imply that equilibrium asset prices are influenced by those high costs, as long as there is an efficient financial service industry with low-cost, reasonably competitive producers.

In considering the preceding technical analysis, one might wonder if there are relevant situations in which the price is observable but trade in the asset cannot take place? One common class of real-world instances is characterized as follows: consider an insurance company that has guaranteed the financial performance of the liabilities of a privately held opaque institution with a mark-to-market portfolio of assets. The market value of that portfolio (corresponding to V in the analysis here) is provided to the guarantor on a continuous basis, but the portfolio itself cannot be traded by the guarantor to hedge its exposure because it does not know the assets held within the portfolio. Elsewhere (Merton, 1997a), I have developed a model using an alternative approach of incentive-contracting combined with the derivative-security technology to analyze the problem of contract guarantees for an opaque institution. It is nevertheless the case that discontinuous tradability of an asset is often accompanied by discontinuous observations of its price. And so, the combination of the two warrants attention. Hence, I comete this section with consideration of how to modify the valuation formula if the price of the underlying asset V is not continuously observable.

Suppose that in the example adopted in this section, the price of the underlying asset is observed at $t = 0$ and then again at the maturity of the derivative contract, $t = T$. In between, there is neither direct observation nor inferential information from payouts on the asset. Hence, $D_1(V, t) = 0$, and the derivative security has no payouts or interim "stopping points" prior to maturity [as specified in (11) and (12)] contingent on $V(t)$. It is however known that the dynamics of V are as described in Assumption 2 with a covariance structure with available traded assets sufficiently well specified to construct the V-Fund according to (5). Define the random variable $X(t) \equiv V(t)/S(t)$, the cumulative *proportional* tracking error, with $X(0) = 1$. By applying Ito's lemma, one has from (8), (9), and (23) that the dynamics for X can be written as

$$dX = \theta X db. \tag{24}$$

It follows from (24) that the distribution for $X(t)$, conditional on $X(0) = 1$, is lognormal with the expected value of $X(t)$ equal to 1 and the variance of $\ln[X(t)]$ equal to $\theta^2 t$. The partial differential equation for F, corresponding to (10), that determines the hedging strategy uses as its independent variable the best estimate of $V(t)$, which is $S(t)$, and it is written as

$$0 = \tfrac{1}{2}\delta^2 S^2 F_{11}[S, t] + rSF_1[S, t] - rF[S, t] + F_2[S, t], \tag{25}$$

subject to the terminal-time boundary condition that for $S(T-) = S$,

$$F[S, T] = E\{h(SX)\}, \tag{26}$$

where h is as defined in (13), X is a log-normally distributed random variable with $E\{X\} = 1$ and variance of $\ln[X]$ equal to $\theta^2 T$ and $E\{\ \}$ is the expectation operator over the distribution of X.

Condition (26) reflects the fact that for all $t < T$, the best estimate of $V(t)$ is $S(t)$. However, at $t = T$, $V(T)$ is revealed and the value of S "jumps" by the total cumulative tracking error of $X(T)$ from its value S at $t = T-$ to $S(T) = V(T)$. The effect of the underlying asset price not being observable is perhaps well illustrated by comparing the solution for the European-type call option with the classic Black-Scholes solution given here in (1). The solution to (25) and (26) with $h(V) = \max[0, V - L]$ is given by, for $0 < t < T$,

$$\begin{aligned} F[S, t]) &= SN(u) - L\exp(-r[T - t]) \\ &\quad \times N(u - \sqrt{\gamma}), \end{aligned} \tag{27}$$

where $u = (\ln[S/L] + r[T - t] + \gamma/2)/\sqrt{\gamma}$, $\gamma = \delta^2(T - t) + \theta^2 T$, *and* $N(\ )$ *is the cumulative density function for the standard normal distribution.*

By inspection of (1) and (27), the key difference in the option-pricing formula with and without continuous observation of the underlying asset price is that the variance over the remaining life of the option does not go to zero as t approaches T, because of the "jump" event at the expiration date corresponding to the cumulative effect of tracking error.

This section has explored conditions under which the Black-Scholes option-pricing model can be validly applied to the pricing of assets with derivative-security-like structures, even when the underlying asset-equivalent is neither continuously traded nor continuously observable. A fuller analysis of this question would certainly take account of the additional tracking error that obtains as a consequence of imperfect dynamic trading of the V-Fund portfolio, along the lines of Bertsimas et al. (1997). However, a more accurate assessment of the real-world impact should also take into account other risk-management tools that intermediaries have to reduce tracking error. For instance, as developed in analytical detail in Merton (1992 pp. 450–57), intermediaries need only use dynamic trading to hedge their *net* derivative-security exposures to various underlying assets. For a real-world intermediary with a large book of various derivative products, netting, which in effect extends the capability for hedging to include trading in securities with "nonlinear" payoff structures, can vastly reduce the size and even the frequency of the hedging transactions necessary to achieve an acceptable level of tracking error. Beyond this, as part of their optimal risk management, intermediaries can "shade" their bid and offer prices among their various products to encourage more or less customer activity in different products to help manage their exposures. The limiting case when the net positions of customer exposures leaves the intermediary with no exposure is called a "matched book."

II. Applications of the Option-Pricing Technology

Open the financial section of a major newspaper almost anywhere in the world and you will find pages devoted to reporting the prices of exchange-traded derivative securities, both futures and options. Along with the vast over-the-counter derivatives market, these exchange markets trade options and futures on individual stocks, stock-index and mutual-fund portfolios, on bonds and other fixed-income securities of every maturity, on currencies, and on commodities including agricultural products, metals, crude oil and refined products, natural gas, and even, electricity. The volume of transactions in these markets is often many times larger than the volume in the underlying cash-market assets. Options have traditionally been used in the purchase of real estate and the acquisition of publishing and movie rights. Employee stock options have long been granted to key employees and today represent a significantly growing proportion of total compensation, especially for the more highly paid workers in the United States. In all these markets, the same option-pricing methodology set forth in the preceding section is widely used both to price and to measure the risk exposure from these derivatives (cf., Robert A. Jarrow and Andrew T. Rudd [1983] and Cox and Mark Rubinstein [1985]). However, financial options represent only one of several categories of applications for the option-pricing technology.

In the late 1960's and early 1970's when the basic research leading to the Black-Scholes
model was underway, options were see
rather arcane and specialized financial in
ments. However, both Black and Sch
(1972, 1973) and I (Merton [1970, 19
recognized early on in the research effort
the same approach used to price options c
be applied to a variety of other valuation p
lems. Perhaps the first major developme
this sort was the pricing of corporate liabili
the "right-hand side" of the firm's bal
sheet. This approach to valuation treated
wide array of instruments used to finance fi
such as debentures, convertible bonds, w
rants, preferred stock, and common stock
well as a variety of hybrid securities) as de
ative securities with their contractual pay
ultimately dependent on the value of the o
all firm. In contrast to the standard fragme
valuation methods of the time, it provide
unified theory for pricing these liabilities.
cause application of the pricing methodol
does not require a history of trading in the
ticular instrument to be evaluated, it was
suited for pricing new types of financial
curities issued by corporations in an inno
ing environment. Applications to corpo
finance along this line developed rapidly.[1

"Option-like" structures were soon see
be lurking everywhere; thus there came an
plosion of research in applying option-pri
which still continues. Indeed, I could no
full justice to the list of contributions a
mulated over the past 25 years even if this
tire paper were devoted to that endea
Fortunately, a major effort to do just th
underway and the results will soon be a
able (Jin et al., 1998). The authors have
erously shared their findings with me. And
I can convey here some sense of the bre
of applications and be necessarily incomp
without harm.

The put option is a basic option which g
its owner the right to sell the underlying a
at a specified ("exercise") price on or be
a given ("expiration") date. When purch
in conjunction with ownership of the un

---

[10] See Merton (1992 pp. 423–27) for an extensiv
of references. See also Gregory D. Hawkins (1982
Michael J. Brennan and Eduardo S. Schwartz (1985a
the early empirical testing by E. Philip Jones et al. (1

lying asset, it is functionally equivalent to an insurance policy that protects its owner against economic loss from a decline in the asset's value below the exercise price for any reason, where the term of the insurance policy corresponds to the expiration date. Hence, option-pricing theory can be applied to value insurance contracts. An early insurance application of the Black-Scholes model was to the pricing of loan guarantees and deposit insurance (cf., Merton, 1977a). A contract that insures against losses in value caused by default on promised payments on a contract in effect is equivalent to a put option on the contract with an exercise price equal to the value of the contract if it were default free. Loan and other contract guarantees, collectively called credit derivatives, are ubiquitous in the private sector. Indeed, whenever a debt instrument is purchased in which there is any chance that the promised payments will not be made, the purchaser is not only lending money but also in effect issuing a loan guarantee as a form of self-insurance. Another private-sector application of options analysis is in the valuation of catastrophic-insurance reinsurance contracts and bonds.[11] Dual funds and exotic options provide various financial-insurance and minimum-return-guarantee products.[12]

Almost surely, the largest issuer of such guarantees are governments. In the United States, the Office of the Management of the Budget is required by law to value those guarantees. The option model has been applied to assess deposit insurance, pension insurance, guarantees of student loans and home mortgages, and loans to small businesses and some large ones as well.[13] The application to government activities goes beyond just providing guarantees. The model has been used to determine the cost of other subsidies including farm-price supports and through-put guarantees for pipelines.[14] It has been applied to value licenses issued with limiting quotas such as for taxis or fisheries or the right to pollute and to value the government's right to change those quotas.[15] Government sanctions patents. The decision whether to spend the resources to acquire a patent depends on the value of the patent which can be framed as an option-pricing problem. Indeed, even on something that is not currently commercial, one may acquire the patent for its "option value," should economic conditions change in an unexpected way.[16] James L. Paddock et al. (1988) show that option value can be a significant proportion of the total valuation of government-granted offshore drilling rights, especially when current and expected future economic conditions would not support development of the fields. Option-pricing analysis quantifies the government's economic decision whether to build roads in less-populated areas depending on whether it has the policy option to abandon rural roads if they are not used enough.[17]

Various legal and tax issues involving policy and behavior have been addressed using the option model. Among them is the valuation of plaintiffs' litigation options, bankruptcy laws including limited-liability provisions, tax delinquency on real estate and other property as an option to abandon or recover the property by paying the arrears, tax evasion, and valuing the tax "timing" option for the capital-gains tax in a circumstance when only realization of losses and gains on investments triggers a taxable event.[18]

---

[11] Cf., Alan Kraus and Ross (1982), Neil A. Doherty and James R. Garven (1986), J. David Cummins (1988), Cummins and Hélyette Geman (1995), and Scott E. Harrington et al. (1995).

[12] Brennan and Schwartz (1976), Jonathan J. Ingersoll, Jr. (1976), M. Barry Goldman et al. (1979), Mary Ann Gatto et al. (1980), and René M. Stulz (1982). In an early real-world application, Myron Scholes and I developed the first options-strategy mutual fund in the United States, Money Market/Options Investments, Inc., in February 1976. The strategy, which invested 90 percent of its assets in money-market instruments and 10 percent in a diversified portfolio of stock call options, provided equity exposure on the upside with a guaranteed "floor" on the value of the portfolio. The return patterns from this and similar "floor" strategies were later published in Merton et al. (1978, 1982).

[13] Howard B. Sosin (1980), Baldwin et al. (1983), Donald F. Cunningham and Patric H. Hendershott (1984), Alan J. Marcus (1987), Merton and Zvi Bodie (1992), Bodie (1996), Ashoka Mody (1996), and Robert S. Neal (1996).

[14] Scott P. Mason and Merton (1985), Calum G. Turvey and Vincent Amanor-Boadu (1989), and Taehoon Kang and B. Wade Brorsen (1995).

[15] James E. Anderson (1987) and Jonathan M. Karpoff (1989).

[16] Lenos Trigeorgis (1993).

[17] Cathy A. Hamlett and C. Phillip Baumel (1990).

In a recent preliminary study, the options structure has been employed to help model the decision of whether the Social Security fund should invest in equities (Kenneth Smetters, 1997). As can be seen in the option formula of the preceding section, the value of an option depends on the volatility of the underlying asset. The Federal Reserve uses as one of its indicators of investor uncertainty about the future course of interest rates, the "implied" volatility derived from option prices on government bonds.[19] In his last paper, published after his death, Black (1995) applies options theory to model the process for the interest rates that govern the dynamics of government bond prices. In another area involving central-bank concerns, Perold (1995) shows how the introduction of various types of derivatives contracts has helped reduce potential systemic-risk problems in the payment system from settlement exposures. The Black-Scholes model can be used to value the "free credit option" implicitly offered to participants, in addition to "float," in markets with other than instantaneous settlement periods. See also Paul H. Kupiec and Patricia A. White (1996). The prospective application of derivative-security technology to enhance central-bank stabilization policies in both interest rates and currencies is discussed in Merton (1995, 1997b).

In an application involving government activities far removed from sophisticated and relatively efficient financial markets, options analysis has been used to provide new insights into optimal government planning policies in developing countries. A view held by some in development economics about the optimal educational policy for less-developed countries is that once the expected future needs for labor-force composition are determined, the optimal education policy should be to pursue targeted training of the specific skills forecast and in the quantities needed. The alternative of providing either more general education and training in multiple skills or training in skills not expected to be used is seen as a "luxury" that poorer, developing countries could not afford. It, of course, was understood, that forecasts of future labor-training needs were not precise. Nevertheless, the basic prescription formally treated them as if they were. In Samantha J. Merton (1992), the question is revisited, this time with an explicit recognition of the uncertainty about future labor requirements embedded in the model. The analysis shows that the value of having the option to change the skill mix and skill type of the labor force over a relatively short period of time can exceed the increased cost in terms of longer education periods or less-deep training in any one skill. The Black-Scholes model is used to quantify that trade-off. In a different context of the private sector in a developed country, the same technique could be used to assess the cost-benefit trade-off for a company to pay a higher wage for a labor force with additional skills not expected to be used in return for the flexibility to employ those skills if the unexpected happens.

The discussion of labor education and training decisions and litigation and taxes leads naturally into the subject of human capital and household decision-making. The individual decision as to how much vocational education to acquire can be formulated as an option-valuation problem in which the optimal exercise conditions reflect when to stop training and start working.[20] In the classic labor-leisure trade-off, one whose job provides the flexibility to increase or decrease the number of hours worked, and hence his total compensation, on relatively short notice, has a valuable option relative to those whose available work hours are fixed.[21] Wage and pension-plan "floors" that provide for a minimum compensation, and even tenure for university professors (John G. McDonald, 1974), have an option-like structure. Other options commonly a part of house-old finance are: the commitment by an institution to provide a mortgage to the house buyer, if he chooses to get one; the prepayment right, after he takes the mortgage, that gives the homeowner the right to renegotiate the interest rate paid to the lender if rates fall;[22] a car lease which gives the customer the right, but not the obligation, to purchase the car at a prespecified price at the end of the lease.[23] Health-care insurance contains varying degrees of flexibility, a major one being whether the consumer agrees in advance to use only a prespecified set of doctors and hospitals ("HMO plan") or he retains the right to choose an "out-of-plan" doctor or hospital ("point-of-service" plan). In the consumer making the decision on which to take and the health insurer assessing the relative cost of providing the two plans, each solves an option-pricing problem as to the value of that flexibility.[24] Much the same structure of valuation occurs in choosing between "pay-per-view" and "flat-fee" payment for cable-television services.

---

[18] George M. Constantinides and Ingersoll (1984), Brendan O'Flaherty (1990), William J. Blanton (1995), Paul G. Mahoney (1995), and Charles T. Terry (1995).

[19] Sylvia Nasar (1992). See Bodie and Merton (1995) for an overview article on implied volatility as an example of the informational role of asset and option prices.

[20] Uri Dothan and Joseph Williams (1981).

[21] Bodie et al. (1992).

Many of the preceding option-pricing applications do not involve financial instruments. The family of such applications is called "real" options. The most developed area for real-option application is investment decisions by firms.[25] However, real-options analysis has also been applied to real-estate investment and development decisions.[26] The common element for using option-pricing here is the same as in the preceding examples: the future is uncertain (if it were not, there would be no need to create options because we know now what we will do later) and in an uncertain environment, having the flexibility to decide what to do after some of that uncertainty is resolved definitely has value. Option-pricing theory provides the means for assessing that value.

The major categories of options within project-investment valuations are: the option to initiate or expand; the option to abandon or contract; and the option to wait, slow down, or speed up development. There are "growth" options which involve creating excess capacity as an option to expand and research and development as creating the opportunity to produce new products and even new businesses, but not the obligation to do so if they are not economically viable.[27]

A few examples: For real-world application of the options technology in valuing product development in the pharmaceutical industry, see Nichols (1994). In the generation of electric power, the power plant can be constructed to use a single fuel such as oil or natural gas or it can be built to operate on either. The value of that option is the ability to use the least-cost, available fuel at each point in time and the cost of that optionality is manifest in both the higher cost of construction and less-efficient energy conversion than with the corresponding specialized equipment. A third example described in Timothy A. Luehrman (1992) comes from the entertainment industry and involves the decision about making a sequel to a movie; the choices are: either to produce both the original movie and its sequel at the same time, or wait and produce the sequel after the success or failure of the original is known. One does not have to be a movie-production expert to guess that the incremental cost of producing the sequel is going to be less if the first path is followed. While this is done, more typically the latter is chosen, especially with higher-budget films. The economic reason is that the second approach provides the option not to make the sequel (if, for example, the original is not a success). If the producer knew (almost certainly) that the sequel will be produced, then the option value of waiting for more information is small and the cost of doing the sequel separately is likely to exceed the benefit. Hence, once again, we see that the amount of uncertainty is critical to the decision, and the option-pricing model provides the means for quantifying the cost/benefit trade-off. As a last example, Baldwin and Clark (1999) develop a model for designing complex production systems focused around the concept of modularity. They exemplify their central theme with several industrial examples which include computer and automobile production. Modularity in production provides options. In assessing the value of modularity for production, they employ an option-pricing type of methodology, where complexity in the production system is comparable to uncertainty in the financial one.[28]

---

[22] Kenneth B. Dunn and John J. McConnell (1981) and Brennan and Schwartz (1985b).

[23] Stephen E. Miller (1995).

[24] James A. Hayes et al. (1993) and Frank T. Magiera and Robert A. McLean (1996).

[25] Mason and Merton (1985), Robert L. McDonald and Daniel R. Siegel (1985), Saman Majd and Robert S. Pindyck (1987), Alexander J. Triantis and James E. Hodder (1990), Avinash K. Dixit and Pindyck (1994), Nancy A. Nichols (1994), Trigeorgis (1996), and Keith J. Leslie and Max P. Michaels (1997).

[26] V. Kerry Smith (1984), Raymond Chiang et al. (1986), David Geltner and William C. Wheaton (1989), Joseph T. Williams (1991), and F. Christian Zinkhan (1991).

[27] W. Carl Kester (1984), Robyn McLaughlin and Robert A. Taggart (1992), and Terrance W. Faulkner (1996).

In each of these real-option examples as with a number of the other applications discussed in this section, the underlying "asset" is rarely traded in anything approximating a continuous market and its price is therefore not continuously observable either. For that reason, this paper, manifestly focused on applications, devotes so much space to the technical section on extending the Black-Scholes option-pricing framework to include nontradability and nonobservability.

III. Future Directions of Applications

As I suggested at the outset, innovation is a central force driving the financial system toward greater economic efficiency with considerable economic benefit having accrued from the changes since the time that the option-pricing papers were published. Indeed, much financial research and broad-based practitioner experience developed over that period have led to vast improvements in our understanding of how to apply the new financial technologies to manage risk. Moreover, we have seen how wide ranging are the applications of our technology for pricing and measuring the risk of derivatives. Nevertheless, there still remains an intense uneasiness among managers, regulators, politicians, the press, and the public over these new derivative-security activities and their perceived risks to financial institutions. And this seems to be the case even though the huge financial disruptions, such as the savings-and-loan debacle of the 1980's in the United States and the current financial crises in Asia and some emerging markets, appear to be the consequence of the more traditional risks taken by institutions such as commercial, real-estate, and less-developed-country lending, loan guarantees, and equity-share holdings.

One conjecture attributes this uneasiness to the frequently cited instances of individual costly events that are alleged to be associated with derivatives, such as the failure of Barings Bank, Proctor and Gamble's losses on complex interest rate contracts, the financial distress of Orange County, and so forth. Perhaps.[29] But, as already noted, derivatives are ubiquitous in the financial world and thus, they are likely to be present in any financial circumstance, whether or not their use has anything causal to do with the resulting financial outcomes. However, even if all these allegations were valid, the sheer fact that we are able to associate individual names with these occurrences instead of mere numbers ("XYZ company" instead of "475–500 thrifts" as the relevant descriptor) would suggest that these are relatively isolated events—unfortunate pathologies rather than indicators of systemic flaws. In contrast, the physiology of this financial technology, that is, how it works when it works as it should, is not the subject of daily reports from around the globe but is essentially taken for granted.

An alternative or supplementary conjecture about the sources of the collective anxiety over derivatives holds that they are a part of a wider implementation of financial innovations which have required major changes in the basic intutional hierarchy and in the infrastructure support it. As a result, the knowledge base now required to manage and oversee financial institutions differs greatly from the traditional training and experience of many financial managers and government regulators. Experiential changes of this sort are threatening. It is difficult to deal with change that is exogenous to our traditional knowledge base and framework and thus comes to seem beyond our control. Decreased understanding of the new environment can create a sense of greater risk even when the objective level of risk in the system remains unchanged or is actually reduced. If so, we should start to deal with the problem now since the knowledge gap may widen if the current pace of financial innovation, as some anticipate, accelerates into the twenty-first century. Moreover, greater complexity of products and the need for more rapid decision-making will probably increase the reliance on models, which in turn implies a growing place for elements of mathematical and computational maturity in the knowledge base of managers. Dealing with this knowledge gap offers considerable challenge to private institutions and government as well as considerable opportunity to schools of management and engineering and to university departments of economics and mathematics.

---

[28] See also Hua He and Pindyck (1992). On an entirely different application, Kester's (1984) analysis of whether to develop products in parallel or sequentially could be applied to the evaluation of alternative strategies for funding basic scientific research: is it better to support N different research approaches simultaneously or just to support one or two and then use the resulting outcomes to sequence future research approaches? See also Merton (1992 p. 426).

[29] Merton H. Miller (1997) provides a cogent analysis refuting many of the specific-case allegations of derivatives misuse.

There are two essentially different frames of reference for trying to analyze and understand changes in the financial system. One perspective takes as given the existing institutional structure of financial service providers, whether governmental or private sector, and examines what can be done to make those institutions perform their particular financial services more efficiently and profitably. An alternative to this traditional institutional perspective—and the one I favor—is the functional perspective, which takes as given the economic functions served by the financial system and examines what is the best institutional structure to perform those functions.[30] The basic functions of a financial system are essentially the same in all economies, which makes them far more stable, across time and across geopolitical borders, than the identity and structure of the institutions performing them. Thus, a functional perspective offers a more robust frame of reference than an institutional one, especially in a rapidly changing financial environment. It is difficult to use institutions as the conceptual "anchor" for analyzing the evolving financial system when the institutional structure is itself changing significantly, as has been the case for the past two decades and as appears likely to continue well into the future. In contrast, in the functional perspective, institutional change is endogenous, and may therefore prove especially useful in predicting the future direction of financial innovation, changes in financial markets and intermediaries, and regulatory design.[31]

The successful private-sector and governmental financial service providers and overseers in the impending future will be those who can address the disruptive aspects of innovation in financial technology while still fully exploiting its efficiency benefits. What types of research and training will be needed to manage financial institutions? The view of the future here as elsewhere in the economic sphere is clouded with significant uncertainties. With this in mind, I nevertheless try my hand at a few thoughts on the direction of change for product and service demands by users of the financial system and the implications of those changes for applications of mathematical financial modeling.

---

[30] For elaboration on the functional perspective, see Merton (1993, 1995), Dwight B. Crane et al. (1995), and Bodie and Merton (1998).

[31] During the last 25 years, finance theory has been a good predictor of future changes in finance practice. That is, when theory seems to suggest that something "should be there" and it isn't, practice has evolved so that it is. The "pure" securities developed by Kenneth J. Arrow (1953) that so clearly explain the theoretical function of financial instruments in risk bearing were nowhere to be found in the real world until the broad development of the options and derivative-security markets. It is now routine for financial engineers to disaggregate the cash flows of various securities into their elemental Arrow-security component parts and then to reaggregate them to create securities with new patterns of cash flows. For the relation between options and Arrow securities and the application of the Black-Scholes model to the synthesis and pricing of Arrow securities, see Ross (1976b), Rolf W. Banz and Miller (1978), Breeden and Robert H. Litzenberger (1978), Darrell J. Duffie and Chi-fu Huang (1986), and Merton (1992 pp. 443–50).

The household sector of users in the more fully developed financial systems has experienced a secular trend of disaggregation in financial services. Some see this trend continuing with existing products such as mutual funds being transported into technologically less-developed systems. Perhaps so, especially in the more immediate future, with the widespread growth of relatively inexpensive Internet accessibility. However, deep and wide-ranging disaggregation has left households with the responsibility for making important and technically complex micro financial decisions involving risk (such as detailed asset allocation and estimates of the optimal level of life-cycle saving for retirement)—decisions that they had not had to make in the past, are not trained to make in the present, and are unlikely to execute efficiently even with attempts at education in the future. The low-cost availability of the Internet does not solve the "principal-agent" problem with respect to financial advice dispensed by an agent. That is why I believe that the trend will shift toward more integrated financial products and services, which are easier to understand and more tailored toward individual profiles. Those products and services will include not only the traditional attempt to achieve an efficient risk-return trade-off for the tangible-wealth portfolio but will also integrate human-capital considerations, hedging, and income and estate tax planning into the asset-allocation decisions. Beyond the advisory role, financial service providers will undertake a role of principal to create financial instruments that eliminate "short-fall" or "basis" risk for households with respect to targeted financial goals such as tuition for children's higher education and desired consumption-smoothing throughout the life cycle (e.g., preserving the household's standard of living in retirement; cf., Franco Modigliani, 1986). The creation of such customized financial instruments will be made economically feasible by the derivative-security pricing technology that permits the construction of custom products at "assembly-line" levels of cost. Paradoxically, making the products more user-friendly and simpler to understand for customers will create conside
more complexity for the producers of
products. Hence, financial-engineering
tivity and the technological and transact
bases to implement that creativity, reliably
cost-effectively, are likely to become a ce
competitive element in the industry. Th
sulting complexity will require more elab
and highly quantitative risk-management
tems within financial service firms and a
allel need for more sophisticated approa
to government oversight. Neither of these
be achieved without greater reliance on ma
matical financial modeling, which in turn
be feasible only with continued improvem
in the sophistication and accuracy of finan
models.

Nonfinancial firms currently use deriva
securities and other contractual agreemen
hedge interest rate, currency, commodity,
even equity price risks. With improved lo
cost technology and learning-curve exp
ence, this practice is likely to expand. E
tually, this alternative to equity capital
cushion for risk could lead to a major cha
of corporate structures as more firms use he
ing to substitute for equity capital, the
moving from publicly traded shares to clo
held private shares.

The preceding section provides example
current applications of the options technol
to corporate project evaluation: the evalua
of research-and-development projects in ph
maceuticals and the value of flexibility in
decision about sequel production in the mo
industry. The big potential shift in the fut
however, is from tactical applications of de
atives to strategic ones.[32] For example, a
pothetical oil company with crude oil rese
and gasoline and heating-oil distribution
no refining capability could complete the
tical integration of the firm by using cont
tual agreements instead of physical acquisi
of a refinery. Thus, by entering into contra
that call for the delivery of crude oil by
firm on one date in return for receiving a

---

[32] See Kester (1984), Stewart C. Myers (1984),
Edward H. Bowman and Dileep Hurry (1993) on the
plication of option-pricing theory to the evaluation of
tegic decisions.

of refined petroleum products at a prespecified later date, the firm in effect creates a synthetic refinery. Real-world strategic examples in natural gas and electricity are described in Harvard Business School case studies, "Enron Gas Services" (1994) and "Tennessee Valley Authority: Option Purchase Agreements" (1996), by Peter Tufano. There is some evidence that these new financial technologies may even lead to a revisiting of the industrial-organization model for these industries.

It is no coincidence that the early strategic applications are in energy- and power-generation industries that need long-term planning horizons and have major fixed-cost components on a large scale with considerable uncertainty. Since energy and power generation are fundamental in every economy, this use for derivatives offers mainline applications in both developed and developing countries. Eventually, such use of derivatives may become standard tools for implementing strategic objectives.

A major requirement for the efficient broad-based application of these contracting technologies in both the household and nonfinancial-firm sectors will be to find effective organizational structures for ensuring contract performance, which includes global clarification and revisions of the treatment of such contractual agreements in bankruptcy. The need for assurances on contract performance is likely to stimulate further development of the financial-guarantee business for financial institutions. Such institutions will have to improve the efficiency of collateral management further as assurance for performance. As we have seen, one early application of the option-pricing model focuses directly on the valuation and risk-exposure measurement of financial guarantees.

A consequence of all this prospective technological change will be the need for greater analytical understanding of valuation and risk management by users, producers, and regulators of derivative securities. Furthermore, improvements in efficiency from derivative products will not be effectively realized without concurrent changes in the financial "infrastructure"—the institutional interfaces between intermediaries and financial markets, regulatory practices, organization of trading, clearing, settlement, other back-office facilities, and management-information systems. To perform its functions as both user and overseer of the financial system, government will need to innovate and make use of derivative-security technology in the provision of risk-accounting standards, designing monetary and fiscal policies, implementing stabilization programs, and overseeing financial-system regulation.

In summary, in the distant past, applications of mathematical models had only limited and sidestream effects on finance practice. But in the last quarter century since the publication of the Black-Scholes option-pricing theory, such models have become mainstream to practitioners in financial institutions and markets around the world. The option-pricing model has played an active role in that transformation. It is safe to say that mathematical models will play an indispensable role in the functioning of the global financial system.

Even this brief discourse on the application to finance practice of mathematical models in general and the option-pricing model in particular would be negligently incomplete without a strong word of caution about their use. At times we can lose sight of the ultimate purpose of the models when their mathematics become too interesting. The mathematics of financial models can be applied precisely, but the models are not at all precise in their application to the complex real world. Their accuracy as a useful approximation to that world varies significantly across time and place. The models should be applied in practice only tentatively, with careful assessment of their limitations in each application.

REFERENCES

Anderson, James E. "Quotas as Options: Optimality and Quota License Pricing Under Uncertainty." *Journal of International Economics*, August 1987, *23*(1/2), pp. 21–39.

Arrow, Kenneth J. "Le Rôle des Valeurs Boursières pour la Répartition la Meilleure des Risques." *Econometrie*, Colloques Internationaux du Centre National de la Recherche Scientifique, 1953, *11*, pp. 41–47.

Bachelier, Louis. "Théorie de la Spéculation." Ph.D. dissertation, l'Ecole Normale Supérieure, 1900. (English translation in Paul H. Cootner, ed., *The random character of stock market prices*. Cambridge, MA: MIT Press, 1964, pp. 17–78.)

Baldwin, Carliss. "Pricing Convertible Preferred Stock According to the Rational Option Pricing Theory." B.S. dissertation, Massachusetts Institute of Technology, 1972.

Baldwin, Carliss and Clark, Kim. *Design rules: The power of modularity*. Cambridge, MA: MIT Press, 1999 (forthcoming).

Baldwin, Carliss; Lessard, Donald and Mason, Scott P. "Budgetary Time Bombs: Controlling Government Loan Guarantees." *Canadian Public Policy*, 1983, *9*, pp. 338–46.

Banz, Rolf W. and Miller, Merton H. "Prices for State-contingent Claims: Some Estimates and Applications." *Journal of Business*, October 1978, *51*(4), pp. 653–72.

Bernstein, Peter L. *Capital ideas: The improbable origins of modern Wall Street*. New York: Free Press, 1992.

Bertsimas, Dimitris; Kogan, Leonid and Lo, Andrew W. "Pricing and Hedging Derivative Securities in Incomplete Markets: An ϵ-Arbitrage Approach." Massachusetts Institute of Technology Working Paper #LFE-1027-97, June 1997.

Black, Fischer. "How We Came Up With the Option Formula." *Journal of Portfolio Management*, Winter 1989, *15*(2), pp. 4–8.

———. "Interest Rates as Options." *Journal of Finance*, December 1995, *50*(5), pp. 1371–76.

Black, Fischer and Scholes, Myron S. "The Valuation of Option Contracts and a Test of Market Efficiency." *Journal of Finance*, May 1972, *27*(2), pp. 399–418.

———. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy*, May–June 1973, *81*(3), pp. 637–54.

Blanton, William J. "Reducing the Value of Plaintiff's Litigation Option in Federal Court: *Daubert v. Merrell Dow Pharmaceuticals, Inc.*" *George Mason Law Review*, Spring 1995, *2*, pp. 159–222.

Bodie, Zvi. "What the Pension Benefit Guaran Corporation Can Learn from the Feder Savings-and-Loan Insurance Corporation *Journal of Financial Services Research*, Jan uary 1996, *10*(1), pp. 83–100.

Bodie, Zvi and Merton, Robert C. "The Informational Role of Asset Prices: The Case of Implied Volatility," in Dwight B. Crane, Kenneth A. Froot, Scott P. Mason, André F. Perold, Robert C. Merton, Zvi Bodie, Eri R. Sirri, and Peter Tufano, eds., *The global financial system: A functional perspective*. Boston, MA: Harvard Business School Press, 1995, pp. 197–224.

———. *Finance*. Upper Saddle River, NJ: Prentice Hall, 1998.

Bodie, Zvi, Merton, Robert C. and Samuelson, William F. "Labor Supply Flexibility and Portfolio Choice in a Life-Cycle Model." *Journal of Economic Dynamics and Control*, July–October 1992, *16*(3–4), pp. 427–49.

Bowman, Edward H. and Hurry, Dileep. "Strategy Through the Option Lens: An Integrated View of Resource Investments and the Incremental-Choice Process." *Academy of Management Review*, October 1993, *18*(4), pp. 760–82.

Breeden, Douglas T. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities." *Journal of Financial Economics*, September 1979, *7*(3) pp. 265–96.

———. "Futures Markets and Commodity Options: Hedging and Optimality in Incomplete Markets." *Journal of Economic Theory*, April 1984, *32*(2), pp. 275–300.

Breeden, Douglas T. and Litzenberger, Robert H. "Prices of State-contingent Claims Implicit in Option Prices." *Journal of Business*, October 1978, *51*(4), pp. 621–51.

Brennan, Michael J. and Schwartz, Eduardo S. "The Pricing of Equity-Linked Life Insurance Policies with an Asset Value Guarantee." *Journal of Financial Economics*, June 1976, *3*(3), pp. 195–213.

———. "Evaluating Natural Resource Investments." *Journal of Business*, April 1985a, *58*(2), pp. 135–57.

———. "Determinants of GNMA Mortgage Prices." *Journal of the American Real Es*

P-11027

B. C. ROY MEMORIAL LIBRARY
INDIAN INSTITUTE OF MA... CALCUTTA

ate & Urban Economics Association, Fall 1985b, *13*(3), pp. 209–28.

ang, Raymond; Lai, Tsong-Yue and Ling, Da- id C. "Retail Leasehold Interests: A Contingent Claim Analysis." *Journal of the American Real Estate & Urban Economics Association*, Summer 1986, *14*(2), pp. 216–29.

nstantinides, George M. and Ingersoll, Jonathan E., Jr. "Optimal Bond Trading with Personal Taxes." *Journal of Financial Economics*, September, 1984, *13*(3), pp. 299–336.

ox, John C. and Ross, Stephen A. "The Valuation of Options for Alternative Stochastic Processes." *Journal of Financial Economics*, January/March 1976, *3*(1/2), pp. 145–66.

ox, John C. and Rubinstein, Mark. *Options markets*. Englewood Cliffs, NJ: Prentice Hall, 1985.

rane, Dwight B.; Froot, Kenneth A.; Mason, Scott P.; Perold, André F.; Merton, Robert C.; Bodie, Zvi; Sirri, Erik R. and Tufano, Peter. *The global financial system: A functional perspective*. Boston, MA: Harvard Business School Press, 1995.

ummins, J. David. "Risk-Based Premiums for Insurance Guarantee Funds." *Journal of Finance*, September 1988, *43*(4), pp. 823–89.

ummins, J. David and Geman, Hélyette. "Pricing Catastrophe Insurance Futures and Call Spreads: An Arbitrage Approach." *Journal of Fixed Income*, March 1995, *4*(4) pp. 46–57.

unningham, Donald F. and Hendershott, Patric H. "Pricing FHA Mortgage Default Insurance." *Housing Finance Review*, December 1984, *3*(4), pp. 373–92.

avis, Mark H. A. "Option Pricing in Incomplete Markets," in M.A.H. Dempster and S. Pliska, eds., *Mathematics of derivative securities*. Cambridge: Cambridge University Press, 1997, pp. 216–26.

ixit, Avinash K. and Pindyck, Robert S. *Investment under uncertainty*. Princeton, NJ: Princeton University Press, 1994.

oherty, Neil A. and Garven, James R. "Price Regulation in Property-Liability Insurance: A Contingent-Claims Approach." *Journal of Finance*, December 1986, *41*(5), pp. 1031–50.

Dothan, Uri and Williams, Joseph. "Education as an Option." *Journal of Business*, January 1981, *54*(1), pp. 117–39.

Duffie, Darrell J. and Huang, Chi-fu. "Implementing Arrow-Debreu Equilibria by Continuous Trading of a Few Long-Lived Securities." *Econometria*, November 1986, *53*(6), pp. 1337–56.

Dunn, Kenneth B. and McConnell, John J. "Valuation of GNMA Mortgage-Backed Securities." *Journal of Finance*, June 1981, *36*(3), pp. 599–616.

Faulkner, Terrance W. "Applying 'Options Thinking' to R & D Valuation." *Research-Technology Management*, May–June 1996, *39*(3), pp. 50–56.

Figlewski, Stephen. "Options Arbitrage in Imperfect Markets." *Journal of Finance*, December 1989, *44*(5), pp. 1289–1311.

Föllmer, Hans and Sondermann, Dieter. "Hedging of Non-Redundant Contingent-Claims," in Werner Hildenbrand and Andreau Mas-Colell, eds., *Contributions to mathematical economics, in honor of Gérard Debreu*. Amsterdam: North-Holland, 1986, pp. 205–23.

Gatto, Mary Ann; Geske, Robert; Litzenberger, Robert H. and Sosin, Howard B. "Mutual Fund Insurance." *Journal of Financial Economics*, September 1980, *8*(3), pp. 283–317.

Geltner, David and Wheaton, William C. "On the Use of the Financial Option Price Model to Value and Explain Vacant Urban Land." *Journal of the American Real Estate & Urban Economics Association*, Summer 1989, *17*(2), pp. 142–58.

Goldman, M. Barry; Sosin, Howard B. and Shepp, Lawrence A. "On Contingent Claims that Insure Ex-Post Optimal Stock Market Timing." *Journal of Finance*, May 1979, *34*(2), pp. 401–13.

Hamlett, Cathy A. and Baumel, C. Phillip. "Rural Road Abandonment: Policy Criteria and Empirical Analysis." *American Journal of Agricultural Economics*, February 1990, *72*(1) pp. 114–20.

Harrington, Scott E.; Mann, Steven V. and Niehaus, Greg. "Insurer Capital Structure Decisions and the Viability of Insurance Derivatives." *Journal of Risk and Insurance*, September 1995, *62*(3), pp. 483–508.

Hawkins, Gregory D. "An Analysis of Revolving Credit Agreements." *Journal of Financial Economics*, March 1982, *10*(1), pp. 59–82.

Hayes, James A.; Cole, Joseph B. and Meiselman, David I. "Health Insurance Derivatives: The Newest Application of Modern Financial Risk Management." *Business Economics*, April 1993, *28*(2), pp. 36–40.

He, Hua and Pindyck, Robert, S. "Investments in Flexible Production Capacity." *Journal of Economic Dynamics and Control*, July–October 1992, *16*(3, 4), pp. 575–99.

Ingersoll, Jonathan E., Jr. "A Theoretical Model and Empirical Investigation of the Dual Purpose Funds: An Application of Contingent-Claims Analysis." *Journal of Financial Economics*, January–March 1976, *3*(1/2), pp. 82–123.

Itô, Kiyoshi. *Kiyoshi Itô selected papers*. New York: Springer-Verlag, 1987.

Jarrow, Robert A. and Rudd, Andrew T. *Option pricing*. Homewood, IL: Irwin, 1983.

Jin, Li; Kogan, Leonid; Lim, Terence; Taylor, Jonathan and Lo, Andrew W. "The Derivatives Sourcebook: A Bibliography of Applications of the Black-Scholes/Merton Option-Pricing Model." Working paper, Massachusetts Institute of Technology, 1998 (forthcoming).

Jones, E. Philip; Mason, Scott P.; Rosenfeld, Eric E. and Fisher, Lawrence. "Contingent Claims Analysis of Corporate Capital Structures: An Empirical Investigation." *Journal of Finance*, July 1984 *39*(3), pp. 611–25.

Kang, Taehoon and Brorsen, B. Wade. "Valuing Target Price Support Programs with Average Option Pricing." *American Journal of Agricultural Economics*, February 1995, *77*(1), pp. 106–18.

Karpoff, Jonathan M. "Characteristics of Limited Entry Fisheries and the Option Component of Entry Licenses." *Land Economics*, November 1989, *65*(4), pp. 386–93.

Kester, W. Carl. "Today's Options for Tomorrow's Growth." *Harvard Business Review*, March–April 1984, *62*(2), pp. 153–60.

Kraus, Alan and Ross, Stephen A. "The Determination of Fair Profits for the Property-Liability Insurance Firm." *Journal of Finance*, September 1982, *37*(4), pp. 1015–28.

Kupiec, Paul H. and White, A. Patricia. "Regulatory Competition and the Efficiency of Alternative Derivative Product Margining Systems." *Finance and Economics Discussion Series*, Board of Governors of the Federal Reserve System, Washington, DC, February 1996.

Leonard, Roger J. "An Empirical Examination of a New General Equilibrium Model for Warrant Pricing." M.S. thesis, Massachusetts Institute of Technology, September 1971.

Leslie, Keith J. and Michaels, Max P. "The Real Power of Real Options." *McKinsey Quarterly*, 1997, (3), pp. 4–22.

Long, John B. "Discussion." *Journal of Finance*, May 1974, *29*(2), pp. 485–88.

Luehrman, Timothy A. "Arundel Partners: The Sequel Project." Harvard Business School Case #9-292-140, 1992.

Magiera, Frank T. and McLean, Robert A. "Strategic Options in Capital Budgeting and Program Selection Under Fee-for-Service and Managed Care." *Health Care Management Review*, Fall 1996, *21*(4), pp. 7–17.

Mahoney, Paul G. "Contract Remedies and Options Pricing." *Journal of Legal Studies*, January 1995, *24*(1), pp. 139–63.

Majd, Saman and Pindyck, Robert S. "Time to Build, Option Value, and Investment Decisions." *Journal of Financial Economics*, March 1987, *18*(1), pp. 7–28.

Marcus, Alan J. "Corporate Pension Policy and the Value of PBGC Insurance," in Zvi Bodie, John B. Shoven, and David A. Wise, eds., *Issues in pension economics*. Chicago: University of Chicago Press, 1987, pp. 49–76.

Mason, Scott P. and Merton, Robert C. "The Role of Contingent Claims Analysis in Corporate Finance," in Edward Altman and Marti Subrahmanyan, eds., *Recent advances in corporate finance*. Homewood, IL: Irwin, 1985, pp. 7–54.

McDonald, John G. "Faculty Tenure As a Put Option: An Economic Interpretation." *Social Science Quarterly*, September 1974, *55*(2), pp. 362–71.

McDonald, Robert L. and Siegel, Daniel R. "Investment and the Valuation of Firms When

There Is an Option to Shut Down." *International Economic Review*, June 1985, *26*(2), pp. 331–49.

McLaughlin, Robyn and Taggart, Robert A. "The Opportunity Cost of Using Excess Capacity." *Financial Management*, Summer 1992, *21*(2), pp. 12–23.

Merton, Robert C. "Lifetime Portfolio Selection Under Uncertainty: The Continuous-Time Case." *Review of Economics and Statistics*, August 1969, *51*(3), pp. 247–57.

______. "A Dynamic General Equilibrium Model of the Asset Market and its Application to the Pricing of the Capital Structure of the Firm." Massachusetts Institute of Technology Working Paper No. 497-70, 1970.

______. "Optimum Consumption and Portfolio Rules in a Continuous-Time Model." *Journal of Economic Theory*, December 1971, *3*(4), pp. 373–413.

______. "Appendix: Continuous-Time Speculative Processes," in R. H. Day and S. M. Robinson, eds., *Mathematical topics in economic theory and computation*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1972, pp. 34–38.

______. "Theory of Rational Option Pricing." *Bell Journal of Economics and Management Science*, Spring 1973a, *4*(1), pp. 141–83.

______. "An Intertemporal Capital Asset Pricing Model." *Econometrica*, September 1973b, *41*(5), pp. 867–87.

______. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance*, May 1974, *29*(2), pp. 449–70.

______. "Option Pricing When Underlying Stock Returns Are Discontinuous." *Journal of Financial Economics*, January/March 1976a, *3*(1/2), pp. 125–44.

______. "The Impact on Option Pricing of Specification Error in the Underlying Stock Price Returns." *Journal of Finance*, May 1976b, *31*(2), pp. 333–50.

______. "An Analytic Derivation of the Cost of Deposit Insurance and Loan Guarantees: An Application of Modern Option Pricing Theory." *Journal of Banking and Finance*, June 1977a, *1*(1), pp. 3–11.

______. "On the Pricing of Contingent Claims and the Modigliani-Miller Theorem." *Journal of Financial Economics*, November 1977b, *5*(3), pp. 241–49.

______. "Financial Economics," in E. C. Brown and R. M. Solow, eds., *Paul Samuelson and modern economic theory*. New York: McGraw-Hill, 1983, pp. 105–40.

______. *Continuous-time finance*. Cambridge, MA: Blackwell, 1992.

______. "Operation and Regulation in Financial Intermediation: A Functional Perspective," in Peter Englund, ed., *Operation and regulation of financial markets*. Stockholm: Economic Council, 1993, pp. 17–67.

______. "Influence of Mathematical Models in Finance on Practice: Past, Present and Future." *Philosophical Transactions of the Royal Society of London*, June 1994, *347*(1684), pp. 451–63.

______. "Financial Innovation and the Management and Regulation of Financial Institutions." *Journal of Banking and Finance*, June 1995, *19*(3, 4), pp. 461–81.

______. "A Model of Contract Guarantees for Credit-Sensitive, Opaque Financial Intermediaries." *European Finance Review*, 1997a, *1*(1), pp. 1–13.

______. "On the Role of the Wiener Process in Finance Theory and Practice: The Case of Replicating Portfolios," in David Jerison, I. M. Singer, and Daniel W. Stroock, eds., *The legacy of Norbert Wiener: A centennial symposium*. Providence, RI: American Mathematical Society, 1997b, pp. 209–21.

Merton, Robert C. and Bodie, Zvi. "On the Management of Financial Guarantees." *Financial Management*, Winter 1992, *21*(4), pp. 87–109.

Merton, Robert C. and Scholes, Myron S. "Fischer Black." *Journal of Finance*, December 1995, *50*(5), pp. 1359–70.

Merton, Robert C.; Scholes, Myron S. and Gladstein, Mathew L. "The Returns and Risk of Alternative Call Option Portfolio Investment Strategies." *Journal of Business*, April 1978, *51*(2), pp. 183–242.

______. "The Returns and Risks of Alternative Put-Option Portfolio Investment Strategies." *Journal of Business*, January 1982, *55*(1), pp. 1–55.

Merton, Samantha J. "Options Pricing in the Real World of Uncertainties: Educating Towards a Flexible Labor Force." B.A. thesis, Harvard University, March 18, 1992.

Miller, Merton H. *Merton Miller on derivatives*. New York: Wiley, 1997.

Miller, Stephen E. "Economics of Automobile Leasing: The Call Option Value." *Journal of Consumer Affairs*, Summer 1995, *29*(1), pp. 199–218.

Modigliani, Franco. "Life Cycle, Individual Thrift, and the Wealth of Nations." *American Economic Review*, June 1986, *76*(3), pp. 297–313.

Mody, Ashoka. "Valuing and Accounting for Loan Guarantees." *World Bank Research Observer*, 1996, *11*(1), pp. 119–42.

Myers, Stewart C. "Finance Theory and Financial Strategy." *Interfaces*, January–February 1984, *14*(1), pp. 126–37.

Nasar, Sylvia. "For Fed, a New Set of Tea Leaves." *New York Times*, July 5, 1992, p. D1.

Neal, Robert S. "Credit Derivatives: New Financial Instruments for Controlling Credit Risk." *Economic Review*, Second Quarter 1996, *81*(2), pp. 14–27.

Nichols, Nancy A. "Scientific Management at Merck: An Interview with CFO Judy Lewent." *Harvard Business Review*, January–February 1994, *72*(1), pp. 89–99.

O'Flaherty, Brendan. "The Option Value of Tax Delinquency: Theory." *Journal of Urban Economics*, November 1990, *28*(3), pp. 287–317.

Paddock, James L.; Siegel, Daniel R. and Smith, James L. "Option Valuation of Claims on Real Assets: The Case of Offshore Petroleum Leases." *Quarterly Journal of Economics*, August 1988, *103*(3), pp. 479–508.

Perold, André F. "BEA Associates: Enhanced Equity Index Funds." Harvard Business School Case #293-024, 1992.

______. "The Payment System and Derivative Instruments," in Dwight B. Crane, Kenneth A. Froot, Scott P. Mason, André F. Perold, Robert C. Merton, Zvi Bodie, Erik R. Sirri, and Peter Tufano, eds., *The global financial system: A financial perspective*. Boston, MA: Harvard Business School Press, 1995, pp. 33–79.

Romano, Marc and Touzi, Nizar. "Contingent Claims and Market Completeness in a Stochastic Volatility Model." *Mathematical Finance*, October 1997, *7*(4), pp. 399–412.

Rosenfeld, Eric. *Stochastic processes of common stock returns: An empirical examination*. Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

Ross, Stephen A. "Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory*, December 1976a, *13*(3), pp. 341–60.

______. "Options and Efficiency." *Quarterly Journal of Economics*, February 1976b, *90*(1), pp. 75–89.

Samuelson, Paul A. "Rational Theory of Warrant Pricing." *Industrial Management Review*, Spring 1965, *6*(2), pp. 13–31.

______. "Mathematics of Speculative Price," in R. H. Day and S. M. Robinson, eds., *Mathematical topics in economic theory and computation*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1972, pp. 1–42

Samuelson, Paul A. and Merton, Robert C. "A Complete Model of Warrant Pricing That Maximizes Utility." *Industrial Management Review*, Winter 1969, *10*(2), pp. 17–46.

Scholes, Myron S. "Taxes and the Pricing of Options." *Journal of Finance*, May 1976, *31*(2), pp. 319–32.

______. "Derivatives in a Dynamic Environment." *American Economic Review*, June 1998, *88*(3), pp. 350–70.

Scholes, Myron S. and Wolfson, Mark A. *Taxes and business strategy: A planning approach*. Englewood Cliffs, NJ: Prentice Hall, 1992.

Smetters, Kenneth. "Investing the Social Security Trust Fund in Equities: An Option Pricing Approach." Technical Paper Series, Macroeconomic Analysis and Tax Analysis Divisions, Washington, DC, 1997.

Smith, Clifford W., Jr. "Option Pricing: A Review." *Journal of Financial Economics*, January/March 1976, *3*(1/2), pp. 3–51.

Smith, V. Kerry. "A Bound for Option Value." *Land Economics*, August 1984, *60*(3), pp. 292–96.

Sosin, Howard B. "On the Valuation of Federal Loan Guarantees to Corporations." *Journal of Finance*, December 1980, *35*(5), pp. 1209–21.

Stratonovich, R. L. *Conditional Markov processes and their application to the theory of optimal control*. New York: American Elsevier, 1968.

Stulz, René M. "Options on the Minimum or the Maximum of Two Risky Assets: Analysis and Applications." *Journal of Financial Economics*, July 1982, *10*(2), pp. 161–85.

Terry, Charles T. "Option Pricing Theory and the Economic Incentive Analysis of Nonrecourse Acquisition Liabilities." *American Journal of Tax Policy*, Fall 1995, *12*(2), pp. 273–397.

Triantis, Alexander J. and Hodder, James E. "Valuing Flexibility as a Complex Option." *Journal of Finance*, June 1990, *45*(2), pp. 549–65.

Trigeorgis, Lenos. "Real Options and Interactions with Financial Flexibility." *Financial Management*, Autumn 1993, *22*(3), pp. 202–24.

———. *Real options*. Cambridge, MA: MIT Press, 1996.

Tufano, Peter. "Enron Gas Services." Harvard Business School Case #294-051, 1994.

———. "Tennessee Valley Authority: Option Purchase Agreements." Harvard Business School Case #296-038, 1996.

Turvey, Calum G. and Amanor-Boadu, Vincent. "Evaluating Premiums for a Farm Income Insurance Policy." *Canadian Journal of Agricultural Economics*, July 1989, *37*(2), pp. 233–47.

Wiggins, James B. "Option Values Under Stochastic Volatility: Theory and Empirical Estimates." *Journal of Financial Economics*, December 1987, *19*(2), pp. 351–72.

Williams, Joseph T. "Real Estate Development as an Option." *Journal of Real Estate Finance and Economics*, June 1991, *4*(2), pp. 191–208.

Zinkhan, F. Christian. "Option-Pricing and Timberland's Land-Use Conversion Option." *Land Economics*, August 1991, *67*(3), pp. 317–25.

Derivatives in a Dynamic Environment[†]

By Myron S. Scholes[*]

I. Introduction

The trading of financial derivatives on organized exchanges has exploded since the early 1970's. The trading of off-exchange financial derivatives on the so-called "over-the-counter" or "OTC" market has exploded since the mid-1980's. Academic and applied research on financial derivatives, which was initiated by the Black-Scholes and Merton option-pricing research in the late 1960's and early 1970's, also has exploded. As a result, three industries have blossomed: an exchange industry in derivatives, an OTC industry in synthetic products, and an academic industry in derivative research, populated by scientists in and out of academic institutions. The academic industry has seen a growth of research and course offerings in economics and business schools in the areas of options, futures, risk management, financial engineering, and—by marrying institutional and derivative modeling—a richer approach to financial intermediation and innovation under uncertainty. In addition, business schools and economics departments have competition from new research and courses in mathematics departments and engineering schools in the mathematics of derivative pricing and alternative stochastic processes, and in law schools in understanding contracting under uncertainty.

The academic industry has produced myriad innovative research papers following the fundamental insights of Fischer Black and Scholes (1973) and Robert C. Merton (1973). The Chicago Board of Trade and Chicago Mercantile Exchange initiated the exchange industry by developing financial options and futures contracts on securities; they have spawned the growth of many new derivative exchanges around the world. The first was the Chicago Board of Trade's sponsorship of the Chicago Board Options Exchange (CBOE) in 1973. Moreover, some exchanges such as the Options Market in Stockholm have transported the technology used to trade derivatives to other markets in Europe and Asia. These exchanges succeed by producing the derivative contracts that add value for individuals and institutions around the world. The OTC industry, which has grown to prominence since the mid-1980's, first in the United States and now in every corner of the globe, is now larger in size than the exchange industry. Financial institutions in the OTC industry offer customized derivative products to meet the specific needs of each of their clients; the exchange industry offers standardized products to reach a richer cross section of demand.

Each industry requires original research to understand the pricing and production costs of the products and financial services that they bring to their clients. Derivative research is quickly transferred to practice; moreover, practice stimulates both academic and applied research. Some of the best research is conducted outside of academic institutions and academics have found a home in each of the three industries. Graduates from mathematics, computer sciences, engineering, and physics compete and cooperate with those trained in financial economics for research and structuring positions in both the exchange and OTC industry.

[†] This article is the lecture Myron S. Scholes delivered in Stockholm, Sweden, December 9, 1997, when he received the Alfred Nobel Memorial Prize in Economic Sciences. The article is copyright © The Nobel Foundation 1997 and is published here with the permission of the Nobel Foundation.

[*] Graduate School of Business, Stanford University, Stanford, CA 94305, and Long-Term Capital Management, L.P., Greenwich, CT 06831. I would like to thank Robert C. Merton for many years of fruitful and exciting discussions on these topics. In addition, I would like to thank my many colleagues at the University of Chicago, MIT, and Stanford University for their ideas and stimulating discussions. Most important has been the support and involvement of Merton H. Miller in my career: I owe him a tremendous debt and cherish his friendship. I miss Fischer Black; I miss his friendship, his insights, and his good humor.

It is difficult to define financial derivatives in a dynamic environment. The purest among us might argue that any security is a derivative if its price dynamics depends on the dynamics of some other underlying asset or assets and price. This broad definition allows not only for what currently exists but also what new derivative instruments will be developed in the future with enhanced understanding and changing production costs. The popular press, however, tends to limit the definition to include financial options, futures, and forward contracts either traded on an exchange or issued in the OTC industry. In the future they may come to be called financial products.

The will of Alfred Nobel states, in part, that the Nobel Prize shall be awarded for an "important discovery or invention." Fischer Black's and my discovery was how to price options and to provide a way to manage risk. Robert C. Merton (1973) developed an important alternative proof of the pricing technology and extended the approach in many directions including how to price options with dividends, how to price options when the interest rate is not constant, and how to apply a more general structure to price many other contingent contracts.

Black and Scholes have over the years been accused of inventing derivatives, at least those derivative products that have been claimed to have had bad economic consequences for their users. It is seldom remembered that these contracts have two sides: if a buyer loses, the ultimate seller might gain. Only if losses cause dead-weight costs is there a net loss to society. It is said that "every successful idea has a thousand fathers, and every failure is an orphan." Over the years we have been granted both distinctions. We did not, however, invent derivatives. Options existed in many financial contracts prior to the Black-Scholes and Merton pricing technologies. Options were noted to have been traded on the Amsterdam Stock Exchange in the late seventeenth century and traded on the Chicago Board of Trade in the early 1930's. As described in Paul H. Cootner (1964), research in option pricing goes back as far as Louis Bachelier's Ph.D. thesis in 1900. Although they are not generally thought of as options, myriad securities and investment decisions have been made in the past and are being made currently that could be evaluated using the Black-Scholes technology. The technology was an invention that facilitated multiple inventions in each of the three industries. What I did not realize at the time of the invention was how the technology would evolve and how it would be used to produce new types of securities with imbedded options at lower cost than could be accomplished prior to the development of the technology. This enhanced the efficiency set of demanders and suppliers of capital, not only in the United States but also around the world.

With the Christmas season approaching at these Nobel award ceremonies in Stockholm, I will invoke the "Past," the "Present," and the "Future," as Charles Dickens did in "A Christmas Carol," to describe the evolution of derivatives in a dynamic environment. I do not mean to draw too fine a parallel, however. That is, I am not implying that through the eyes of Mr. Scrooge the past for derivatives was bright with hope and innocence, and the present is dark and foreboding, and the future, without changing our ways, presents a bleak picture. On the contrary, many in academics and those in practice have often asked for a glimpse of the past; that is, how we developed the technology and the model, and the past gives insights into the present. It was a time of innocence. It was a time of discovery. It is a tradition at these Nobel award ceremony lectures to describe the age of innocence. I am honored and thank the Nobel Prize Committee for this wonderful opportunity do so. I wish that Fischer Black were alive today to share this honor with us.

Twenty-five years is a tender age for the new academic and the new exchange industry. Fifteen years is still a young age for the new OTC industry. The Present, which I date from the late 1980's to current time, shows an industry that has experienced growing pains, and many, including regulators, are worried that it still has not come of age. And, will the Future be bleak? No, there will be failures, but the industries will thrive because derivative instruments will provide progressively lower-cost solutions to investor and entity problems than will competing alternatives. These lower-cost solutions will involve the unbundling and repackaging of coarse financial products into their constituent parts to satisfy client demands. The process will continue to evolve as advances in information technology drive down the cost of providing alternative and more productive solutions.

II. The Past

Although Black (1989) and Peter L. Bernstein (1992) described their versions of the development of the option-pricing technology, this lecture, however, gives me an opportunity to add to their description through my recollection of things past. My formal training at the University of Chicago was in financial economics, statistics, and economics. At Chicago, Merton H. Miller, the 1990 Nobel laureate, and Eugene Fama, a prolific scholar, stimulated my excitement in economics and a new branch of economics, which has come to be called financial economics. I also owe a similar and considerable debt to my fellow classmates at Chicago, most notably, Jack Gould, Michael C. Jensen, and Richard Roll.

The three strands of financial economics that most set the tone for my future research were arbitrage and the notion of substitutes; the efficient markets hypothesis; and the capital asset pricing model.[1] These strands gave a mathematical basis to the models of finance, in a general equilibrium framework. Franco Modigliani and Miller (see Miller [1988] for a retrospective) were making profound breakthroughs that provided a general equilibrium model for corporate finance. Their arbitrage arguments, which demonstrated how a firm's value was independent of how it financed its activities, had a profound effect on the way I analyze and model many problems in economics. The Fama (1965) and Paul A. Samuelson (1965b) efficient-markets hypothesis which states that, in a well-functioning capital market, the dynamics of asset prices are described by a submartingale and that the best estimate of the value of a security is today's price, was revolutionary to economics. Their insights
gave me an important framework to
about the dynamics of asset prices an
markets adjust to "news." It set the sta
empirical testing of how information w
corporated into security prices and gave
vehicle to apply my statistics and con
skills in a financial economics context. F
ample, Jensen (1968) had used the co
of the efficient-markets hypothesis t
—whether mutual funds, which were profe
ally managed and spent considerable su
money to discover undervalued assets,
outperform randomly selected investme
a risk-adjusted basis. Roll (1970) had
the efficiency of the bond market contr
for changes in expected returns impound
bond prices. In my own work, Scholes (1
I used the concepts of efficient market
substitutes (arbitrage) to test the exte
which security prices were influenced b
size of the sales of large blocks of secu
or by changes in the information set. Fo
ing on the work of Harry M. Markowitz (
1959), the William F. Sharpe (1964)
John Lintner (1965) capital asset pri
model provided a general equilibrium m
of asset prices under uncertainty. This be
the fundamental model for measuring the
of a security. It was elegant to condens
required relative rates of return on secu
into a simple reduced-form equation tha
pended only on their "betas," a measu
their relative contribution to the risk of th
timal portfolio.

The capital structure models, the effic
markets hypothesis, and the capital a
pricing model had the common ce
themes of arbitrage and market equilibr
securities with similar economic risk ha
exhibit similar returns to prevent arbit
profits. This principle applies to all sec
ties, whether they are common stock, bo
or hybrid instruments. Through particip
in seminar presentations at the Universit
Chicago, I became generally aware of th
ture of warrants and convertible bonds. I
unaware, however, of the interest in wa
pricing and the research that had been c
ducted at MIT on this topic in the 196
even though I became an assistant profe
at the Sloan School of Management at
in 1968.

---

[1] I am sure that Fischer Black would pick at least the efficient-markets hypothesis and the capital asset pricing model as most influential in the development of his thinking in financial economics.

A032901
B. C. ROY MEMORIAL LIBRARY
INDIAN INSTITUTE OF MANAGEMENT CALCUTTA

I did not meet Robert Merton until the spring of 1969, when he was interviewing for a position at the Sloan School. We began interacting in the fall of 1969. We talked then about his current research on dynamic applications of the capital asset pricing model with changing opportunity sets and my work, at the time, including tests of the capital asset pricing model. We did not talk about warrant pricing even though Fischer Black and I were working on the problem. I guess we did not appreciate that each of us had an interest in this research area.

In the summer of 1968, a research project at Wells Fargo Bank in San Francisco convinced me that the passive management of assets could be a viable contender to actively managed portfolios. In my report to Wells Fargo Bank, I recommended that they initiate passive investment strategies and offer them to their clients, the forerunner to so-called "index funds." That fall, on the suggestion of Michael Jensen, I had lunch with Fischer Black, who he had met because of Jensen's research on mutual funds. Fischer was employed at Arthur D. Little, a consulting firm in Cambridge, MA. We had several other lunches that fall and Fischer suggested that he was thinking of leaving Arthur D. Little to start his own consulting firm. At about the same time, John McQuown at Wells Fargo Bank asked whether I wanted to conduct research that would describe the trade-off of risk and return in the market as a forerunner to introducing passive investment strategies to their clients. Without research, they were not willing to offer index-fund-like products to clients. Being an assistant professor, I was restricted as to the number of days I could consult. I asked Fischer whether he wanted to join forces on the project. It was obvious from our lunch discussions that he had very similar ideas, and was starting a research project with Jensen on measuring risk and return. We joined forces and worked together to test the capital asset pricing model (see Black et al., 1972). We developed the concept of the zero-beta portfolio to test the model. If we could create a zero beta-minimum variance portfolio by buying low beta stocks and selling high beta stocks and achieve realized returns on this portfolio that were significantly different from the interest rate, this would violate the predictions of the original capital asset pricing model.

In the winter of 1969, I agreed to direct the master's thesis of an MIT graduate student who had garnered a time series of warrant and underlying stock prices and wanted to apply the capital asset pricing model to value the warrants. I read all of the articles relating to warrant pricing in Cootners' book of readings on *The Random Character of Stock Prices* (1964). One included paper, by Case Sprenkle and dated 1960, seemed the most relevant to me, but Sprenkle used an exogenously determined discount rate to discount the expected terminal value of the warrant to its present value.

What seemed apparent was that the expected return of the warrant could not be constant for each time period because the risk of the warrant changed with changes in the stock price and with changes in time to maturity. For example, if the warrant was far "in-the-money," that is, the underlying stock price was far above the exercise price, and the warrant was almost sure to be exercised, its price would change almost dollar for dollar with a change in the underlying stock price. The percentage change in the value of the warrant, however, would be greater than the percentage change in the value of the common stock because the warrant was a leveraged instrument. On the other hand, if the warrant was "out-of-the-money," that is, the underlying stock price was less than the exercise price, the warrant price would move far less than dollar for dollar with the price of the common stock (for example, $0.5 for $1 move in the common). The percentage change in the price of the warrant, however, would be even greater than that of the in-the-money warrant.

As a result, the expected return on the warrant could not be constant each period if the beta of the stock was constant each period. I thought about using the capital asset pricing model to establish a zero-beta portfolio of common stock and warrants by selling enough shares of common stock per each warrant held each period to create a zero-beta portfolio. Given I could create a zero-beta portfolio, the expected return on the net investment in this portfolio would be equal to the riskless rate of interest. I knew that I would have to change the number of shares of stock each period to retain my zero-beta portfolio. But, after working on this concept, off and on, I still could not figure out analytically how many shares of stock to sell short to create a zero-beta portfolio.

Fischer and I continued working on tests of the capital asset pricing model and the development of investment products based on the implications of our research throughout 1969. In the summer or early fall of 1969, I discussed with Fischer my earlier experience with warrants, my attempt at creating the zero-beta portfolio, and my inability to determine the changing number of shares needed each period to create the zero-beta portfolio. He described to me his research on warrants and that he was frustrated in his inability to progress further than he had to that time. He showed me a sheet of paper, which described the relation between the return on the warrant and the underlying stock. Following on earlier work by Jack Treynor, Fischer had used a Taylor-series expansion of $w(x, t)$, where "w" is the warrant price, "x" is the current stock price, and "t" is time to maturity to show the relation between the change in the warrant price as a function of the change in the price of the common stock and a decrease in the time to maturity of the option. Ignoring terms of second order with regard to time, over a short period of time, this expansion was:

$$\Delta w(x, t) = w_1\Delta x + w_2\Delta t + \tfrac{1}{2} w_{11}\Delta x^2\Delta t, \tag{1}$$

where Δ is the change symbol, and the subscripts refer to partial derivatives with respect to the first or second arguments.

Not surprisingly, Fischer had used the capital asset pricing model to describe the relation between the expected return on the warrant and the market and the expected return on the common stock and the market. By substituting for the change in the warrant price as a function of changes in the stock price and time in the capital asset pricing relation, it became obvious on how to create a zero-beta portfolio that would have an expected rate of return equal to the interest rate (for we assu constant interest rate).

Consider the returns over a very short p of time on two alternative investment s gies: under strategy (1), we acquire the rant, and enough bonds earning at intere per period, such that our investment in str (1) is the same as in alternative strategy in which we buy w_1 of stock. The follow the investment and the return on these tw ternative strategies:

	Investment		Return
(1)	Buy warrant	w	$w_1\Delta x + w_2\Delta t + \frac{1}{2} w_{11}\Delta$
	bonds	$w_1 x - w$	$r\Delta t(w_1 x - w)$
(2)	Buy stock	$w_1 x$	$w_1\Delta x.$

The investment was constructed to b same in strategy (1) and strategy (2). risk appears to be the same in both strate The only uncertain term is Δx, the chan the stock price, and the total uncertainty to changes in the stock price is the sam both strategies. Δx^2 involves the chan the stock price squared, a form of vari which as Δt becomes small approaches the stock price squared times the inst neous variance of the underlying return the common stock, which is assumed t constant.

Since the risk is the same and the invest is the same under both strategies, to pre arbitrage, the returns must be the same o short period of time. After equating the ret on strategy (1) with the returns on stra (2), and substituting for Δx^2, we find:

$$-rw + w_1 xr + w_2 + \tfrac{1}{2} w_{11} x^2\sigma^2 = 0 \tag{2}$$

This is the Black-Scholes differential e tion. The initial condition for a warrant or option is that $w(x, t^*) = \max(x - c,$ where t^* is the maturity date of the option, c is the exercise price of the option. The required inputs to value the option, other its initial conditions, are r, the interest r and σ^2, the variance rate per unit time on returns on the underlying stock. We were b amazed that the expected rate of return on underlying stock did not appear in the di ential equation.

Although the number of shares needed to create a zero-beta portfolio each period was w_1, it was not obvious to us how to find w_1. The next step in solving the problem was to realize that since the warrant valuation depended only on the variability of returns and not the expected return on the underlying common stock, it was arbitrary what expected return was assumed for the underlying common stock. The same warrant valuation equation would result because we had hedged out the risk of the common stock in establishing the zero-beta portfolio or the alternative replicating portfolio, as above. We assumed that the expected return on the common stock was equal to the interest rate over the next short period of time, or in terms of the capital asset pricing model that the common stock had a zero beta.

With the assumption of constant return and variance of return, the distribution of returns on the underlying stock at expiration of the warrant would be lognormally distributed. We used the Sprenkle formulation to find the terminal value of the warrant using a constant interest rate as the expected rate of return on the stock. But, we wanted the present value of the warrant. The key here was to realize that although the warrant would have greater price variability than the underlying stock, if we assume that the stock had a zero beta, the warrant would have a zero beta. If the warrant had a zero beta each period of time, the warrant had also to return the interest rate, r, each period of time.

If we had decided to value the warrant using the actual expected return on the common stock or, for that matter, any other appreciation rate, the discount rate to value the warrant would depend on time and changes in the stock price. It does not for the zero-beta case. Using Sprenkle's formula with the assumption that the expected return on common stock and the discount rate for the warrant was equal to a constant interest rate, we obtained the Black-Scholes option-pricing formula.

$$(3) \quad w(x, t) = xN(d_1) - ce^{r(t-t^*)}N(d_2),$$

where $N(d)$ is the cumulative normal density function, c is the exercise price, t^* is the expiration date, and t is the current date, $t^* - t$ is the remaining number of periods in the life of the option, and "e" is the exponential operator. Lastly,

$$(4) \quad d_1 = [\ln x/c + (r + 1/2\sigma^2)(t^* - t)] / \sigma\sqrt{t^* - t}$$

$$d_2 = d_1 - \sigma\sqrt{t^* - t}.$$

We checked the formula against the differential equation and, as we expected, it fit. We were sure that we had the correct formula for valuing call options or warrants, the right to buy an asset at a fixed price, the exercise price, at maturity of the right, its expiration date. With minor adjustments we could value a put option, the right to sell an asset under similar terms.

From the formula, we were finally able to compute w_1, which was equal to $N(d_1)$, the required number of shares to hedge the option. The number of shares will change over time and as the price of the underlying security changes with respect to the exercise price. But, given the assumptions it is a known quantity each time period.

The formulation also suggests the technology necessary to price any contingent claim that depends on an underlying asset's price (or even other state variables) and time. This is so even with differential known payouts, such as dividends on a stock or coupons on a bond, that are not received by the option holder. The technology suggests that what is necessary is to hedge the stochastic term, Δx, to create an alternative investment that is riskless. Merton (1973) formalized these relations.

We had spent a considerable amount of time working to finish up several other papers including our paper on testing the capital asset pricing model. As a result, we finished a draft of the paper sometime in the winter of 1970. We did not know whether our formulation was exact, but intuitively we thought investors could diversify away any residual risk that was left. For larger price changes in the common stock, the hedge position of being long the option and short the appropriate number of shares of stock would tend to make money whether the market went up or down, and would lose money on small changes in the market portfolio. The risk of the position appeared to be independent of market risk. We programmed the model and tried to understand the sensitivity of the price of the option to changes in the stock price, time to maturity, volatility, and the interest rate.

We presented a draft of the paper at a conference on capital market theory sponsored by Wells Fargo Bank in July of 1970. Later that summer, on vacation together, Fischer and I worked out the applications of the option-pricing framework to the pricing of risky debt and other capital structure issues. We viewed the common stock of a company with debt in its capital structure as a call option. The equity holders have an option to buy back the firm from the debt holders by paying off the face amount of the debt at its maturity. The equity holders will not buy back the firm from the debt holders at the maturity of the debt if the face amount of the debt is less than the value of the firm's assets; they will turn the remaining assets of the firm over to the debt holders. For us, it was exciting to realize that the equity of a corporation was an option, and that our framework applied far more broadly than the valuation of warrants or put and call options. The methodology provided a systematic approach that relied on arbitrage to value the capital structures of firms, and to understand how management decisions affect the relative values of debt and equity in the firm's capital structure.

As it turned out, Robert Merton, who had written an earlier paper on the valuation of warrants with Samuelson (see Samuleson and Merton, 1969), following on Samuleson's (1965a) early work on warrant valuation, had expected to attend our session at the Wells Fargo conference, but he overslept and missed the session. In the winter and spring of 1970, Fischer and I searched the academic literature to determine how close others were to our invention. Fischer and I realized that the Samuelson and Merton paper contained an equilibrium model to value warrants but they did not value the warrant continuously. Given that there was friendly rivalry between the two teams, Fischer and I wanted to progress, on our own, as far as we could prior to the conference.

After the conference, in the early fall, Robert Merton and I discussed the Black-Scholes valuation methodology and option-valuation formula. He was not convinced that using the capital asset pricing model framework was sufficient. It seemed to him that, as the interval of adjustment of the hedge became closer to continuous time, there might still be covariance between the hedged position and the market return. Merton and I speculated that one way that there would be no such covariance would be if the return on the option was perfectly correlated with the return on the stock in continuous time, and therefore the hedge was exact. Merton (1973) later proved that the hedged position in continuous time was riskless, and that the replicating portfolio argument was exact. Fischer and I used this derivation in the final version of the paper because it relied on arbitrage and not on any underlying model of capital market equilibrium. However, we still presented the capital asset pricing model derivation of the model, as it provided us with the many insights we used to unlock the puzzles of option pricing.

Merton (1973) then started working on his paper on various aspects of the option formula. He incorporated his alternative proof of the option-pricing model. He also showed that the right to exercise a call option prior to maturity is not valuable for a nondividend paying stock, but valuable for a put option on a similar stock. He also showed how to incorporate changes in interest rates into the valuation methodology, and generalized the formula to handle other state variables.

To our dismay, when we submitted a version of our paper entitled "A Theoretical Valuation Formula for Options, Warrants, and Other Securities" (October 1970) to the *Journal of Political Economy,* it was rejected without review. The *Review of Economics and Statistics* also rejected the paper. Fischer felt that the paper was rejected because he was not an academic; I felt that I was an unknown assistant professor and the paper would not be considered to be broad enough for those academic journals. With the help of Merton Miller and Eugene Fama, who took an interest in the paper and stepped in on our behalf, the *Journal of Political Economy* agreed to consider the paper if we revised it and broadened its applicability. We had planned to publish the corporate-finance-capital structure applicaions in a subsequent paper but we broadened the original paper and showed how corporate liabilities could be viewed as options. The final version of the paper was published in 1973 in the *Journal of Political Economy* under the title "The Pricing of Options and Corporate Liabilities."

A. *The Aftermath*

In Black and Scholes (1972), we tested the option-pricing model using data recorded in a transactions diary of a broker in the over-the-counter option's market provided to us by another master's student at MIT. The diary listed the prices at which he sold options to his clients. Using simple estimates of the volatility, the model generally performed well. It produced profits if one could buy options at diary-market prices if the model indicated a higher value than the market and sell options at diary-market prices if the model indicated a lower value than the market. Each trading day, we assumed that the model could be used to determine the hedge ratio, the delta, and undertook these hedges (assuming that transaction costs were zero). We regressed the daily returns on our hedged portfolio on the market returns. As expected, the portfolio returns were uncorrelated with the market returns, but the model produced substantial and significant abnormal profits. The market appeared to ignore information available in the historical data on estimating volatility.

When we assumed that we could buy the undervalued and sell the overvalued options at model prices and hedge out the underlying stock risk, we incurred significant losses. These losses were incurred because using simple estimates of the volatility ignored information on future volatility that the market was using to price the options. When actual realized volatility over the life of the option was used to compute the model prices, buying undervalued and selling overvalued options at model prices generated returns that were insignificantly different from zero.

It became apparent to us that the transaction costs of dealing in the over-the-counter market were quite large. The dealers would only sell options. As a result, the market in put and call options was quite small. The world was about to change. In the early 1970's, various studies were commissioned to provide an economic justification for a new options exchange. As a result, the Chicago Board Options Exchange was born in 1973 almost simultaneously with the publication of the Black-Scholes and Merton option-pricing papers. The reduction in transaction costs and the transparency of the market were justifications enough for the subsequent success of the options market.

It is ironic that these empirical tests of the Black-Scholes model were published in the *Journal of Finance* in the proceedings volume of the *American Finance Association Meetings*, in May of 1972, a full year prior to the publication of the model itself. Although we did not present it as such, it is ironic that the methodology in the paper is generally the same as that used today by financial entities to manage the risks of their trading positions and to measure the performance of their traders.

B. *Historical Notes: From Theory to Practice*

Both the derivative exchange industry and the derivative academic industry grew significantly from 1973 to 1985. Financial economists started to interact with a broad set of practitioners and this led to a cross-fertilization of ideas among the participants in both industries.

The option-pricing technology was adopted simply because it reduced transaction costs. For without a model, traders could neither price securities with imbedded options with sufficient accuracy to compete against other traders with models, nor could they reduce the risk of their positions to employ their capital efficiently at a low enough cost to compete with other traders. Although it is hard to prove, I do think that the success of the CBOE and other exchanges, in part, can be attributable to option-pricing models. As traders became familiar with these models, bid-offer spreads narrowed. As traders became more familiar with risk-management techniques, they could take on larger position sizes to support the market. With a deeper and more efficient market, investors began to use options to facilitate their own investment strategies.

In those formative years, notable extensions and additions to the basic framework include important contributions by Merton (1974, 1992a), Black (1975, 1976), John C. Cox and Stephen A. Ross (1976), Ross (1976), Mark Rubinstein (1976), Scholes (1976), Michael Parkinson (1977), Oldrich A. Vasicek (1977), Rolf W. Banz and Miller (1978), Douglas T. Breeden and Robert H. Litzenberger (1978), William Margrabe (1978), Scott F. Richard (1978), Sharpe (1978), Michael J. Brennan and Eduardo S. Schwartz (1979), Cox et al. (1979), Robert Geske (1979), and Michael Harrison and David M. Kreps (1979).

In 1971, Fischer Black left Boston to become a professor at the University of Chicago. In 1972, Robert Merton and I became consultants to Donaldson, Lufkin and Jenrette (DLJ) to build mathematical models to price so-called "down-and-out options" and to build their options technology to price call options in the event of a launch of the CBOE. Leo Pomerance, the head of the DLJ options group, was an options trader from the old school; he traded OTC options using intuition and experience without regard to a formal model. He later became the first chairman of the CBOE. At DLJ we forged a marriage of the old-time trader types, with their mental set, with young mathematical modeling types, with their model assumptions, to add value for the firm.

The spread of the option-valuation technology was rapid once the CBOE launched its first contracts on listed securities. Initially, many of the older market-wise traders rejected using a model to price options. And, initially prices were not in line with prices predicted from using the Black-Scholes model adjusting for dividends and using relatively simple estimates of volatility to price the options. This left an opportunity for younger model-based traders to step in and profit from price discrepancies in the market. They used the model to price options and to determine the appropriate hedge ratios to reduce the risk of their positions. Generally, retaining these risks had zero present value because traders had little expertise in forecasting stock returns. By so doing they could undertake larger positions and enhance their profits by concentrating on risks that could add to their profits.

In Dan Galai (1975), Galai, one of my Ph.D. students at the University of Chicago, where I had returned in 1973, tested the pricing of options on the CBOE in the first year of its existence using the model with simple historical estimates of the volatility. He found that the profits on trading options, that is, buying undervalued contracts and selling overvalued contracts each day to maintain a neutral risk position, were even greater than those found in our original tests. His strategies could achieve greater profits by reducing positions if the prices of options return to model values prior to the expiration of the contract. Transaction costs could have reduced actual trading profits for other than the option dealers.

By the end of the first year of trading options, it was no longer possible to use simple estimates of the historical volatility to spot opportunities in the market. Many of the clearing firms, which financed the positions of the option traders, used the model and the hedge ratios to determine the net risk of each of these traders. Fischer had started a service to provide option prices and the share-equivalent positions (hedge ratios) on each of the options traded on the various exchanges. He used a more sophisticated estimate of volatility to price the options. He combined historical estimates adjusted for changes in stock market levels, with the volatility implied by the prices of options. As is true even to this day, as the market-price level of securities increases relative to a previous level over a relatively short period of time, the volatility of stocks tends to fall. This result is due, in part, to a reduction in the leverage of the underlying equities. Given the rate of interest, the price of the common stock, the dividend yields, the exercise price, and the maturity date of the option, the model can be used to infer the implied volatility that the market is using to price the option. In fact, even today, options are described in terms of implied volatility. Traders are asked whether they want to buy or sell volatility. Although his volatility estimates held some cache for a while, advanced computer technology made Fischer's pricing sheets obsolete after a few years.

In fact, Texas Instruments marketed a handheld calculator in 1974 that gave the Black-Scholes model values and hedge ratios.

When I asked them for royalties, they replied that our work was in the public domain; when I asked, at least, for a calculator, they suggested that I buy one. I never did. Robert Merton and I continued to consult for DLJ. Working with Mathew L. Gladstein, we decided that the time might be appropriate to provide investors with a fund that protected their downside risk but allowed for some upside participation in the performance of the stock market. To achieve this goal, we decided that the assets of the fund would be held in two parts: in any six-month period, 90 percent of the assets would be held in U.S. Treasury bills and 10 percent of the assets would be used to buy a diversified portfolio of call options. In several papers, Merton et al. (1978, 1982) simulated the performance characteristics of such a strategy using the underlying stocks of the options that were traded in the various markets. The return characteristics were as predicted by the theory: losses were truncated and gains were less than a direct investment in the underlying stocks. The returns on the strategy were nonlinearly related to the market. We always stressed the role of options as insurance. In early 1976, we attempted to launch Money Market/Options Investments under the auspices of Phoenix Investment Counsel of Boston. Unfortunately, the fund raised only a small amount of money. Our simulation results indicated that fully covered investment strategies, that is, a sale of an option on a position in the underlying stock (for example, long 100 shares of IBM and short a call option to buy 100 shares of IBM), would provide returns of only the premium received on the option approximately 60 percent of the time. Call-option holders would call their stock away if the stock price were above the exercise price on expiration of the contract and would not exercise call options if the stock price were below the exercise price at expiration. This strategy produces higher current income but with the possibility of capital losses just like a high-yield bond. The expected return, however, was less than the expected return on the money market-options strategy. At about this time other investment companies marketed the fully covered strategy, which we thought exhibited inferior return characteristics for most investors, and naturally, to our dismay, were quite successful because of the promises of higher income (but at the unadvertised expense of expected capital losses).

III. The Present

The past 20 years has seen a transformation of the entire financial services industry, first in the United States, and now around the world. During the 1970's and 1980's, regulations divided the activities of financial institutions into separate market segments. In the United States, commercial banks handled deposits and made commercial loans; investment banks were involved in mergers, acquisitions, and underwriting; brokerage companies sold stocks and bonds; savings and loans, along with banks, initiated and held mortgages; and, insurance companies sold life, and property and casualty insurance products. Many of the regulations were directed at preserving the profitability of these institutions by restricting competition, mainly at the expense of the users of these services. Each institution had a product focus; for example, deposits, or life insurance, or commercial loans. No financial company served a broad range of its clients' financial needs.

As happens at times, it is not possible for regulators to protect the profitability of the industries they regulate. In the United States, mutual funds competed with banks in providing deposit services after banks were not permitted to pay market interest rates on deposits in the early 1970's. The growth of institutional investors managing pension funds and mutual funds forced the abolition of fixed commission rates to trade securities in the United States and around the world. The larger brokerage firms evolved to compete against the banks and the savings and loan associations in packaging and repackaging mortgages to broaden the extent of the market. Banks started to compete with investment banks and brokerage and insurance companies in financing commercial real estate and financing highly leveraged mergers and acquisitions called "Leveraged Buyouts." It is nearly impossible to maintain regulations that restrict activities in one industry when new competitors not subject to costly regulations are attacking the profitable businesses of that industry.

The driving force behind today's tidal wave of financial innovation has been the reduction in the cost of computer and communications technology. This lower-cost technology has led to a globalization of the product and financial markets. Corporate and institutional needs have become more complex. Investors are demanding more services. Technology brings new competitors to the market who can offer similar and expanded services at lower cost than existing competitors can. The growth of lower-cost providers of brokerage services such as Schwab, Fidelity, and Internet brokers are examples.

Financial service firms today must decide which clients to serve, determine those clients' needs, and then decide which products and services add value for their clients as well as their own shareholders. Firms that were quite similar 15 years ago have become very different today. For instance, J.P. Morgan, a wholesale bank, today differs far more from Citicorp, a retail bank, than it did in the early 1980's. Conversely, firms that were quite different from each other 15 years ago have become quite similar today. It is hard to distinguish the activities of today's J.P. Morgan from Union Bank of Switzerland (UBS), from Goldman Sachs, or Merrill Lynch. Meanwhile, American Insurance Group (AIG) and Travelers, both insurance companies, these days offer a number of services that are similar to those offered by Goldman Sachs or UBS. Investment banks no longer merely structure and advise in transactions but instead have moved to a more packaged, integrated, convenient financial-solutions approach, directed at solving the complex problems of their clients around the world.

The many advances in financial theory have enabled financial services firms to meet those complex needs more effectively and at lower cost than was possible previously. The marriage of business school and economics department graduates with engineers, mathematicians, physicists, and computer scientists has lead to more efficient and lower-cost financial engineering solutions to client problems.

To date the major growth in the use of
rivatives has been fueled by trends towa
curitization and the increased understandi
the role that derivatives can play in the un
dling, packaging, and transferring of risks
longer do financial service firms only sel
same products they buy from clients. Ins
they break the products down into their c
ponent parts and either sell the parts or rec
bine them into new and hybrid custom-tail
financial instruments. And, this unbund
and repackaging is only in the begin
stages of evolution.

With information asymmetries between
ents and their financial service provide
would be prohibitively expensive for a cl
to develop close relations with many finan
service firms. To be productive, the finan
service firm must learn the needs of its cli
and understand their businesses. As a resul
can be expert to only a select client list. Cli
find it inefficient to "shop" widely for s
financial service firms. Creating custo
tailored solutions strengthens relations
tween the financial services firm and
clients. It would be too costly for each cl
to replicate the specialized expertise requ
to engineer financial theory solutions; such
ent would be underutilized most of the ti
It would be analogous to every corpora
maintaining an entire full-service law firm
its premises. Notwithstanding many reg
tors' fears, it is not likely that all financial s
vice firms will disappear if left to compet
the global arena. Product standardization
erode profits more quickly than in the past
cause more diverse entities, such as Gene
Electric, Enron, or accounting firms, can co
pete in providing financial services using
nancial technology. New competition
enter various markets from global competit
Although inefficient financial service fi
will disappear more quickly than in the p
their clients will obtain more value-enhanc
and less-costly services from the remaining
nancial service firms. Financial products
becoming so specialized that, for the m
part, it would be prohibitively expensive
trade them in organized markets.

Financial service firms have become
leaders in using derivatives in their ri
management programs. Using informat

and option-pricing technology, financial services firms can not only value their commitments, such as guarantees and other derivative contracts, but also are moving to understand the sensitivities of their holdings to various market factors. They can decide what risks to transfer and what risks to retain.

Tables 1 and 2 show the growth of derivative contracts trading since 1986 in both the exchange industry and in the OTC industry. The cells of the tables contain the "notional principal amounts outstanding" for various categories of derivative contracts. For example, the face amount of stock market index options (including call and put options) at the end of 1986 stood at $37.8 billion and by the end of 1996 grew to $380.2 billion.[2]

These tables indicate that the OTC market in derivatives has grown much faster than the exchange market in the last ten years. In 1995, turnover on the major derivative exchanges around the world actually declined while OTC activity rose by 40 percent. The Bank for International Settlements estimated that outstanding OTC contracts exceeded $47.5 trillion in early 1995, much greater than the numbers reported in Table 2. To put these values into perspective, the value of all outstanding debt in Europe, Japan, and North America totaled $25.8 trillion in 1995.

The growth of the OTC market will continue to outstrip the growth of the exchange market because the clients of the financial service firms need assistance to structure their financial programs. The current growth path is to provide more client-focused structured solutions to problems. Clients likely would find it less expensive to execute a program through their financial service firm than to execute it themselves in the exchange market. This is even more likely if the positions must be adjusted frequently to hedge risks.

Moreover, the relative growth of the OTC market is overstated because the exchange markets require that entities post margin on contracts each day. Futures contracts are settled at the end of each day. Forward contracts, such as swaps and options written by OTC firms such as caps and floors, are not settled each day. The product offerings are different. The financial service firms and their counterparts rely on each other's credit. Most financial service firms can post collateral on OTC contracts, which is very similar to settling the contracts as in the case of posting margin on an exchange. These entities will use either the financial futures and options markets or the OTC markets. They will use the industry that provides services at lower cost. Many entities, however, are not indifferent to posting collateral. In particular, many OTC swap and option contracts have a financial entity and a corporation as counterparts because the corporation is willing to pay the financial entity to post margin for it in the futures or options market. That is, the financial services firm enters into a swap with a corporate entity, which does not post collateral, and the financial service firm hedges its market risk by entering into an offsetting swap with another financial service firm or by using the exchange-derivatives industry. In either case, the financial service firm posts collateral or margin on its transactions. It is the lower-cost producer of margin services. Financial service firms have the capacity and the personnel to undertake the pricing of credit risk and can handle these transactions; many corporate entities currently do not.

To the extent that collateral is not posted on the obligation, one consequence of the these transactions is that the financial service firm and the corporation are exposed to each other's credit risk. The corporation buying a call option on an underlying debt instrument from the financial service firm has credit risk to the extent of the value of the call option. The financial service firm can fail to honor its obligation to deliver the underlying debt instrument. The financial service firm holding a put option, issued by the corporation, on the

---

[2] See International Monetary Fund (1997) for the source of these statistics. The notional amount outstanding is not an economic measure of the size of the market. The notional amount of a swap or an option is the amount on which the contract is based. It is not the value that the security would trade at in the market. For example, if a call option to buy $100,000 of a major-market index trades at $5,000, the notional amount is recorded as $100,000 while the economic value is only $5,000. The economic value of swaps and options might be as low as 2 percent and 5 percent of the notional outstanding amounts of the contracts. These statistics provide estimates of the growth of the derivative market.

TABLE 1—MARKETS FOR SELECTED DERIVATIVE FINANCIAL INSTRUMENTS: NOTIONAL PRINCIPAL AMOUNTS OUTSTANDING: 1986–1996 (IN BILLIONS OF U.S. DOLLARS)

	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Interest rate futures	370.0	487.7	895.4	1,200.8	1,454.5	2,156.7	2,913.0	4,958.7	5,777.6	5,863.4
Interest rate options[a]	146.5	122.6	279.2	387.9	599.5	1,072.6	1,385.4	2,362.4	2,623.6	2,741.8
Currency futures	10.2	14.6	12.1	16.0	17.0	18.3	26.5	34.7	40.1	38.3
Currency options[a]	39.2	59.5	48.0	50.2	56.5	62.9	71.1	75.6	55.6	43.2
Stock market index futures	14.5	17.8	27.1	41.3	69.1	76.0	79.8	110.0	127.3	172.2
Stock market index options[a]	37.8	27.7	42.9	70.7	93.7	132.8	158.6	229.7	238.3	329.3
Total	618.3	729.9	1,304.8	1,766.9	2,290.4	3,519.3	4,634.4	7.771.1	8,862.5	9,188.2
North America	518.1	578.1	951.7	1,155.8	1,268.5	2,151.7	2,694.7	4,358.6	4,819.5	4,849.6
Europe	13.1	13.3	177.7	251.0	461.2	710.1	1,114.3	1,777.9	1,831.7	2,241.6
Asia-Pacific	87.0	138.5	175.4	360.0	560.5	657.0	823.5	1,606.0	2,171.8	1,990.1
Other	0.0	0.0	0.0	0.1	0.2	0.5	1.8	28.7	39.5	106.8

[a] Calls plus puts.

same debt instrument is a creditor of the corporation to the extent of the value of the put option. A swap contract, which states that the corporation will receive a fixed rate of interest on an underlying debt instrument and pay a floating rate of interest, is equivalent to the corporation being long a call option and short a put option.

The other major reason that the OTC industry will continue to grow faster than the exchange industry is that financial service firms and others need only to hedge the remaining factor risks of their portfolio positions, which is a far smaller amount than their gross contracting with their clients. Moreover, depending on the costs, they can hedge either with another financial service firm or in the exchange industry. The financial service firm that hedges factor risks retains the remaining risk, the so-called "basis risk," of its net positions.

Another reason for the growth of the OTC market has been that the outstanding amounts in Table 2 do not necessarily represent net exposures. It might be less expensive for a corporation or a financial service firm to enter into an offsetting derivative contract with another counterparty than it would be to unwind the initial contract. If it does, the contract volu increases but the net exposure falls.

A. *The Present: The Pathologies*

From the perspective of market comment tors, many regulators, and the public, deriv tives tend to top the list of suspects when th stock market turns downward or when entiti announce unanticipated financial losses. Th press, the public, and regulators fear deriv tives, in part, because they are new and, i part, because their growth has appeared to b so explosive over the last ten years. Althoug they vastly overstate the economic exposur notional amounts as high as $45 trillion caus worry. The press and others credit the mark crash of 1987 to portfolio insurance, an a tempt to dynamically replicate the returns o options. Even in this time, market pundit warn that forms of dynamic hedging could fo ter a severe market downturn. Widely publi cized losses attributed to derivative trading i the 1990's include: the leveraged-derivati contracts issued by Bankers Trust to firm such as Proctor and Gamble and Gibson Gree ing Cards (over $150 million in losses); th loss of $1.5 billion by Shell Sekiyu; the loss

TABLE 2—NOTIONAL PRINCIPAL VALUE OF OUTSTANDING INTEREST RATE AND CURRENCY SWAPS OF THE MEMBERS OF THE INTERNATIONAL SWAPS AND DERIVATIVES ASSOCIATION, 1987–JUNE 1996 (IN BILLIONS OF U.S. DOLLARS)

	1987	1988	1989	1990	1991	1992	1993	1994	1995	June 1996
Interest rate swaps										
All counterparties	682.9	1,101.2	1,502.6	2,311.5	3,065.1	3,850.8	6,177.3	8,815.6	12,810.7	15,584.2
Interbank (ISDA member)	206.6	341.3	547.1	909.5	1,342.3	1,880.8	2,967.9	4,533.9	7,100.6	—
Financial Institutions	300.0	421.3	579.2	817.1	985.7	1,061.1	1,715.7	2,144.4	3,435.0	—
Governments[a]	47.6	63.2	76.2	136.9	165.5	242.8	327.1	307.6	500.9	—
Corporations[b]	128.6	168.9	295.2	447.9	571.7	666.2	1,166.6	1,829.8	1,774.2	—
Currency swaps										
All counterparties (adjusted for reporting of both sides)	182.8	319.6	449.1	577.5	807.2	860.4	899.6	914.8	1,197.4	1,294.7
Interest rate options[c]	0.0	327.3	537.3	561.3	577.2	634.5	1,397.6	1,572.8	3,704.5	4,190.1
Total	865.6	1,657.1	2,489.0	3,450.3	4,449.5	5,345.7	8,474.5	11,303.2	17,712.6	21,068.9

[a] Including international institutions.

[b] Including others.

[c] Including caps, collars, floors, and swaptions.

incurred by Orange County investing in inverse floaters; the bankruptcy of Metallgesellschaft and Barings Corporation (both over $1 billion in losses); and many other losses by financial service firms such as UBS, Salomon Brothers, etc. Obviously, many of these losses are overstated because there were gains made by the other side to these contracts: It is only the dead-weight costs to society that result in actual loss. For excellent discussions of the entire range of purported pathologies and an excellent review of the literature addressing these issues, see Miller (1997), who argues that most, if not all, of the "diagnoses" of severe pathologies are misdiagnosed.

Yet, the growth of these industries depicted in Tables 1 and 2 clearly suggests that these instruments have added net value. It is hard to believe such growth could continue for so many years without value being realized by the clients of financial service firms, the shareholders of these firms, and the exchanges. I have argued in Scholes (1995, 1996) that the development of financial infrastructure might lag financial innovation. It is costly to develop controls and firm-wide understanding of new products that are in the prototype phase of their development. Prototypes are built using existing infrastructure. For an innovation as long lasting and profitable as derivatives, the OTC industry, the exchange industry, and the academic industry find it profitable to build the infrastructure necessary to support them. Each of these industries has a vested interest in profiting from adding value that is sustainable, so each will attempt to invest in the cost-effective infrastructure necessary to preserve this value.

This is not to argue that we have seen the end of derivative failures. There will be losses sustained as in many other business activities. In 1997, the governments of many countries in Asia could no longer support the losses of their financial institutions resulting from defaulted commercial and real estate loans. Although many banks had been economically bankrupt without the support of their governments (that is, the value of their equity would have been effectively zero without government-implicit guarantees), the governments allowed these banks to participate in any potential gains and sheltered them against bank runs by promising to pay off their depositors. These options were costly to society, and it will be difficult to prove that they were value enhancing. In part, other countries, through the grants made by the

International Monetary Fund as the lenders of last resort to these and other countries, may have written the put options that supported the activities of the financial institutions of the region and granting these options might have encouraged risk taking and even unprofitable activities.

Moreover, absent government guarantees, some brokerage houses and banks in Japan probably would be bankrupt, because while these entities had promised clients protection against loss on any decline in value of Japanese stocks, they did not hedge their commitments. Clients might have paid for these put options through higher commission rates in Japan. These Japanese financial service entities, however, suffered severe losses when their clients exercised these put rights during the decline in the value of Japanese stocks in the 1990's. The entities currently are not required to value their commitments on a mark-to-the-market basis. As a result, neither regulators nor investors, nor even senior management, could deduce the financial condition of the entities. In all probability, the extent of these losses could have been mitigated if risk-management policies had been put in place.

IV. The Future

The future will be a continuation of the present. Financial innovation will continue at the same, or at even an accelerating, pace because of the insatiable demand for lower-cost, more efficient solutions to client problems. Information and financial technology will continue to expand and so will the circle of understanding of how to use this technology. There is value to investing in education. Financial service firms will expand the use of this technology to manage their own activities. Otherwise, they will have to face mergers with other financial service entities. Although some would like to see derivatives wither in importance, they will not, for they have become essential mechanisms in the tool kit of financial innovation.

Scholes and Mark A. Wolfson (1992) used the concepts of frictions and restrictions to illustrate how tax rules and other regulations affect investor and corporate behavior. As Merton (1992a) argues, the functions of a financial system change far less than institutions. Institutions change because lower-cost solutions that reduce information asymmetries are found to facilitate transactions, to provide funding for large-scale investment projects, to transfer savings across borders and into the future, and to provide more efficient risk-sharing and diversification mechanisms.

Most financial instruments are derivative contracts in one form or another. Black and Scholes (1973) pointed out that the equity holders of a firm with debt in its capital structure have an option to buy back the firm from its debtholders at maturity of the debt. The high-yield bond (the so-called "junk bond") is a riskier debt-option contract than more highly rated corporate debt. Corporate debt and equity contracts are derivative to underlying investments. Other lines of research on so-called "real options" indicate that even the investment decisions of firms are better understood by using an option framework rather than a more conventional present-value-analysis framework.

Standard debt and equity contracts are institutional arrangements or boxes. They provide particular cash flows to investors with their own particular risk and return characteristics. These institutional arrangements survive only because they provide lower-cost solutions than competing alternative arrangements. Competitive opportunities evolve over time with changing frictions and restrictions. Because of information asymmetries and regulatory restrictions, investors might require a higher rate of return to hold these standard-form contracts than contracts (now and in the future) of alternative design but of similar risk. Time will continue to blur the distinctions between debt and equity.

The firm's investment set is generally the composite of coarse bundles of payoffs. Firms issue claims to finance these activities—claims that themselves represent bundles of coarse cash flows. It will become more efficient for financial service firms to offer new derivative securities in various forms to break cash flows into finer gradients that can be tailored to the specific needs of demanders and suppliers of capital. In the process, deadweight costs are mitigated, thereby reducing the cost of capital. The financial service firm can sell the newly created securities or retain them in whole or in part for its own account. It can create new products on its own name or use the OTC or exchange markets to hedge its risks. Given information asymmetries, it will use the lower-cost solution.

A. *Investor Demands*

In recent years, we have witnessed a movement from a limited number of investors holding an undiversified portfolio of their own home-country securities to many more investors holding diversified portfolios domestically and internationally. More and more investors around the world, who have never invested in financial products other than through social promises made by their governments, will become more willing to select from a broad class of "mutual-fund" type offerings. Although the diversity of products has grown, few tools are in place other than in academic circles that allow investors to make informed portfolio-allocation decisions. As Franco Modigliani (1986), the 1985 Nobel laureate, has argued: individuals want to smooth consumption over the life cycle. If it were more cost efficient, investors would want to insure against contingencies, control risks more efficiently, and plan their investments efficiently to meet life-cycle needs. As information and financial technology become more easily available, financial service firms will repackage investments to meet these investor demands, and this will spur financial innovation. The financial service firm will offer products in its own name that promise specific risk and return patterns; these firms will also offer products in the form of mutual funds. The classifications of investment products into stock funds, bond funds, growth, income, etc., will diminish in importance with reduced costs to understand risk, return, and contingent payoffs.

Even today, the boxes that define institutional-fund arrangements have blurred. For example, if a pension fund manager wants to achieve a stock-index fund return, she can invest with an index-fund provider that buys a diversified portfolio of the index-fund stocks. She can achieve the same result, however, in myriad other ways, including using another manager who might claim to have expertise in the bond market and can provide an enhanced return over the index-fund return. To achieve this, the enhanced manager might undertake a complicated strategy. He might hold a portfolio of undervalued corporate and government bonds, hedge the credit risk of the corporate bonds by selling stock short, and hedge the price risk of interest rate movements by using futures or options. He might buy stock-index futures to achieve the systematic risk exposure of the stock-index fund. Given costs, the manager might be able to produce a return in excess of that achieved by holding the index fund directly while the systematic risks of the two offerings would be exactly the same.

The exchange industry will compete with the OTC industry to provide investment products. If the exchange industry can provide efficient margining systems for those investors who can not post collateral in a cost-efficient manner, those products that become standardized will most likely be ideally suited for the exchange industry—it can address a larger set of participants in a lower-cost marketplace. The exchange industry complements the OTC industry; they will grow together.

B. *Corporate Demands for Derivatives*

Finance specialists have puzzled over the reasons why corporations hedge the risks of their cash flows. Under classical finance theory, it is often asked why shareholders of a firm pay it to incur costs to reduce risk when they can diversify on their own account. The firm's managers should act as if the firm is risk neutral. Clifford W. Smith and René M. Stulz (1985) provide three reasons, all tied to the cost of financial distress, why a firm might hedge its cash flows. Because of the convexity of the tax schedule, a firm might issue more debt only after hedging its cash flows to reduce its expected operating losses and the resultant loss of tax benefits. Because of bankruptcy costs associated with high levels of debt, the firm that hedges can use more debt to finance its activities. As in Froot et al. (1993), if the firm can hedge its cash flows, a reduction in the probability of financial distress reduces the expected costs of financial distress and, as a result, encourages investment in profitable projects that might have been foregone without such hedging. This argument is based on the observation that firms are reluctant to issue equity and, instead, use retained earnings to finance investments before using the debt markets. Also, if a high-debt-to-equity firm were to become financially distressed it would not be possible to issue equity to finance business activities. Because owner-managers in smaller firms might not be able to diversify their holdings, hedging the cash flows of the firm might be a lower-cost alternative than selling off pieces of the firm to outsiders and using the proceeds to diversify. If corporations face these problems, there are financial engineering solutions that might reduce their import.

The corporate use of derivatives is not limited to hedging. Some corporate financial strategies believe that they can outperform other market participants in forecasting the direction of interest rates or commodity prices. Stulz (1996) argues that firms that have such financial acumen can hedge their downside exposures by buying put options. This allows the financial officers of the firm to become more active managers. By buying put options or by reducing systematic risks, they can use more leverage and increase their personal stakes by reducing the costs of financial distress. This tactical use of derivatives probably explains a significant part of the growth of the use of derivatives attributable to financial service firms and corporations, as shown in Tables 1 and 2.

As reported in Stulz (1996), empirical evidence gathered from surveys of corporations indicates that large corporations without debt in their capital structures hedge cash flows more so than smaller corporations. And, those corporations that do hedge lift their hedges from time to time or do not fully hedge their exposures. It is these tactical uses of derivatives, an attempt to "beat" the market using highly leveraged strategies, that have been the cause of most of the reported financial losses. Obviously, the successful tactical users of derivatives are most often absent in press reports. It is unlikely, however, that corporate officials, on average, can outperform other market participants.

Large firms hedge cash flows, in part, to smooth reported financial earnings with the hope those smoother earnings will boost their price-to-earnings ratios. This might be a value-enhancing strategy if market participants can not discern whether the variability in earnings is caused by the firm's taking systematic exposures to market factors or by firm-specific risks.

The Present is still young. The Future will bring many new solutions to solve corporate problems. Many corporations and financial entities still need to learn and evaluate to what extent hedging and risk control can be beneficial to their activities. Smaller firms and product markets are just now becoming familiar with the risk-control aspects of these financial instruments. It may be surprising that in the United States the top eight banks account for 94 percent (almost $19 trillion) of the total outstanding notional amount in the OTC market as of the end of 1996.[3] As of this date, the knowledge base or the financial acumen needed to financial engineer solutions to client problems is highly concentrated.

As in Scholes (1995, 1996), I argue that corporations will use risk-management techniques to reduce their level of equity capital and, as a result of risk-management techniques, some firms that would have gone public will remain private. Equity capital is an expensive form of financing. There are large differences between the knowledge base of insiders and outsiders. Insiders can not fully divulge their plans to outsiders for the fear that competitors will profit from this knowledge, and generally must sell shares at a discount. Moreover, tax and other considerations make the corporate form of undertaking activities in the United States and in other countries very expensive.

Equity is a risk-management device. It is an "all purpose" risk cushion. The more equity a firm has, the more it cushions itself against outcomes that require it to go the capital markets in adverse times or when it might have to divulge its confidential operating plans to outside parties. Hedging, on the other hand, is targeted risk control. Hedging requires more refined knowledge of the firm, and an understanding of the interaction of investment returns and financing alternatives. Moreover, it requires that the firm be able to warrant to others that it will maintain a strategy of hedging its activities to support higher levels of debt. But as the costs to hedge fall relative to the costs of equity, firms will substitute hedging for equity.

---

[3] See International Monetary Fund, 1997.

Moreover, hedging provides ancillary benefits as a measurement tool to help calibrate how the firm is making money. In a diverse, decentralized organization, management information systems might not divulge the true source of profits within the organization; that is, did profits arise because systematic risk exposures produced positive returns, or because the entity possessed superior skills? Standard accounting neither provides risk-management reports which decompose profitability into profits from market forces and profits from managerial efforts (whether the firm is a manufacturing or a financial firm), nor does it describe the sensitivities of the firm's profit and loss to market factors. As more entities use financial engineering skills, the current accounting system will be under considerable pressure for change, as will many of the current forms of regulations and restrictions.

Because of differences in the required knowledge of insiders and outsiders, the growth of the private equities market has reduced the disclosure costs of becoming a public corporation. Private equity allows expert management teams to leverage their activities. Private equity, however, is still an inefficient form of financing compared with potential lower-cost solutions. Ways will be found through financial engineering to provide private entities with the advantages of the public market—risk sharing, liquidity, and pricing signals—while retaining the advantages of the private market—lower disclosure and agency costs. Financial engineering will foster the growth of the private corporation, and convert entities into alternative forms.

Many firms hedge interest rate movements, foreign currency exposures, or commodity price exposures. Firms will learn to use stock-index options or futures to reduce their risk exposures. The firm can reduce the beta of its own stock by hedging stock market risks. Moreover, with this approach, the firm does not have to target risks. It can just hedge its own market risks or other factor risks, or the general stock market risk that affect its stock price. This reduces the economic risks of the firm to firm-specific risks, or residual risks, and reduces the need for equity to cushion adverse outcomes.

I believe that the corporate form we know today will not be long-lived. With more knowledge and a better understanding of the power of financial engineering and of how to reduce asymmetric information costs, the costs of using financial engineering solutions will continue to fall. As more firms learn how to use these solutions, their profits will be enhanced and more investment will follow the increase in demand. Risk management is only a step in the direction of producing synthetic entities.

The firm of the future might be an organizational form far different from those used today. Some entities, such as electricity producers, aircraft manufacturers and users, natural resource producers and users, and financial service firms already are deciding what services to produce and what risks to retain, and what services to rent and what risks to shed, based on their perceived competitive advantage.

Financial service firms are building large capital bases to make markets around the world, and to put into practice specific knowledge to engineer solutions for their clients on a global basis and to create long-lived derivative products for issue in their own names. Their profits are made from modeling and understanding markets and providing value-added solutions for clients. A risk-management system provides information on what risks to keep and what risks to hedge. In addition, it provides a way to reduce information asymmetries between senior management and employees, and to provide the incentive system necessary to align the interests of the employees and the firm's shareholders.

A risk-management system must also address how a financial service firm handles crisis situations. To preserve its franchise, a financial service firm can insure against adverse price movements or unforeseen contingencies by holding working capital as a reserve against adverse liquidity needs. Alternatively, the financial service firm might be able to buy options from the exchange markets or from the OTC markets (for example, lines of credit) at a lower dead-weight cost to insure against extreme price movements that could adversely affect its business. Maybe regulatory bodies, in effect, provide lower-cost insurance.

Because of tax and regulatory costs, financial service firms might find that working capital held in corporate form is too expensive relative to other alternatives. For example, clients of financial service firms hold large quantities of passive wealth in mutual funds, insurance companies, pension funds, and various trusts as investment vehicles for individual savings. Financial service firms and other entities must hold working capital to insure against adverse contingencies. With options and other forms of contingent capital arrangements, it will become possible to mobilize the capital in these client-passive investment vehicles and reduce the dead-weight costs of the current system. This could lower the cost of capital of financial service firms and the cost of providing financial services to their client base.

Once again, information and financial technology will expand to reduce the costs of information asymmetries. Understanding and developing markets for credit derivatives, understanding the implications of contingent capital options under asymmetric information, and understanding what is the most efficient mechanisms to hold capital will change organizational forms, the boxes, and blur the distinctions between debt and equity, corporations and partnerships, and the demanders and suppliers of capital. The evolution of option technology will open up entire new institutional structures.

V. Conclusion

We started in the Past—the age of innocence, and we progressed to the Present—the age of understanding, growth, and maturation. The growth has not been without pains. A considerable amount of additional understanding and development awaits the users of the derivative technology in the Future—the age of excitement. Advances in communications and computing technology will allow for greater reduction in asymmetric information costs. The future growth of innovation using the option-pricing technology will be as great or greater than in the past. Organizational forms will change dramatically in a global environment. The exchange industry will continue to grow; the OTC industry will continue to grow, and the research necessary by academics and practitioners to understand and to foster the evolution will be so great that the academic industry will become more important than in the past. The need for highly trained and skilled practitioners who understand the technology will continue to increase on a global basis. None of these three industries has retained its past form and none will retain its present form. These economic organizational forms always will evolve and respond to the demands of economic agents.

The capital asset pricing model, arbitrage and capital structure models, and the efficient-markets hypothesis introduced me to key finance concepts that were the genesis of our development of the option-pricing technology. In a world of information asymmetries, derivative instruments provide lower-cost solutions to financial contracting problems in a dynamic environment and these lower-cost solutions enhance economic efficiency.

REFERENCES

Bachelier, Louis. "Théorie de la Spéculation." Ph.D. dissertation, l'Ecole Normale Supérieure, 1900. (English translation in Paul H. Cootner, ed., *The random character of stock market prices*. Cambridge, MA: MIT Press, 1964, pp. 17–78.)

Banz, Rolf W. and Miller, Merton H. "Prices for State-contingent Claims: Some Estimates and Applications." *Journal of Business*, October 1978, *51*(4), pp. 653–72.

Bernstein, Peter L. *Capital ideas: The improbable origins of modern Wall Street*. New York: Free Press, 1992.

Black, Fischer. "How We Came Up With the Option Formula." *Journal of Portfolio Management*, Winter 1989, *15*(2), pp. 4–8.

______. "Fact and Fantasy in the Use of Options." *Financial Analysts Journal*, July/August 1975, *31*, pp. 36–41, 61–72.

______. "The Pricing of Commodity Contracts." *Journal of Financial Economics*, January/March 1976, *3*(1/2), pp. 167–79.

Black, Fischer; Jensen Michael C. and Scholes, Myron S. "The Capital Asset Pricing Model: Some Empirical Tests," in Michael C. Jensen, ed., *Studies in the theory of capital markets*. New York: Praeger, 1972, pp. 79–121.

Black, Fischer and Scholes, Myron S. "The Valuation of Option Contracts and a Test of Market Efficiency." *Journal of Finance*, May 1972, *27*(2), pp. 399–417.

______. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy*, May–June 1973, *81*(3), pp. 637–54.

Breeden, Douglas T. and Litzenberger, Robert H. "Prices of State-contingent Claims Implicit in Option Prices." *Journal of Business*, October 1978, *51*(4), pp. 621–51.

Brennan, Michael J. and Schwartz, Eduardo S. "A Continuous Time Approach to the Pricing of Bonds." *Journal of Banking and Finance*, July 1979, *3*(2), pp. 133–55.

Cootner, Paul H., ed. *The random character of stock market prices*. Cambridge, MA: MIT Press, 1964.

Cox, John C. and Ross, Stephen A. "The Valuation of Options for Alternative Stochastic Processes." *Journal of Financial Economics*, January/March 1976, *3*(1/2), pp. 145–66.

Cox, John C.; Ross, Stephen A. and Rubinstein, Mark. "Option Pricing: A Simplified Approach." *Journal of Financial Economics*, September 1979, *7*(3), pp. 229–63.

Fama, Eugene. "The Behavior of Stock Prices." *Journal of Business*, January 1965, *37*(1), pp. 34–105.

Froot, Kenneth A.; Scharfstein, David S. and Stein, Jeremy C. "Risk Management: Coordinating Corporate Investment and Financing Policies." *Journal of Finance*, December 1993, *48*(5), pp. 1629–58.

Galai, Dan. "Pricing of Options and the Efficiency of the Chicago Board Options Exchange." Ph.D. dissertation, University of Chicago, 1975.

Geske, Robert. "The Valuation of Compound Options." *Journal of Financial Economics*, March 1979, *7*(1), pp. 63–81.

Harrison, J. Michael and Kreps, David M. "Martingales and Arbitrage in Multiperiod Securities Markets." *Journal of Economic Theory*, June 1979, *2*(3), pp. 381–408.

International Monetary Fund. *International capital markets, developments, prospects and key policy issues*. Washington, DC: International Monetary Fund, 1997.

Jensen, Michael C. "The Performance of Mutual Funds in the Period 1945–64." *Journal of Finance*, May 1968, *23*(2), pp. 389–416.

Lintner, John. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics*, February 1965, *47*(1), pp. 768–83.

Margrabe, William. "The Value of an Option to Exchange One Asset for Another." *Journal of Finance*, March 1978, *33*(1), pp. 177–86.

Markowitz, Harry M. "Portfolio Selection." *Journal of Finance*, March 1952, *7*(1), pp. 77–91.

______. *Portfolio selection: Efficient diversification of investment*. New York: Wiley, 1959.

Merton, Robert C. "Theory of Rational Option Pricing." *Bell Journal of Economics and Management Sciences*, Spring 1973, *4*(1), pp. 141–83.

______. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance*, May 1974, *29*(2), pp. 449–70.

______. "Financial Innovation and Economic Performance." *Journal of Applied Corporate Finance*, Winter 1992a, *4*(4), pp. 12–22.

______. *Continuous-time finance*. Cambridge, MA: Blackwell, 1992b.

Merton, Robert C.; Scholes, Myron S. and Gladstein, Mathew L. "The Returns and Risk of Alternative Call Option Portfolio Strategies." *Journal of Business*, April 1978, *51*(2), pp. 183–242.

______. "The Returns and Risks of Alternative Put-Option Portfolio Investment Strategies." *Journal of Business*, January 1982, *55*(1), pp. 1–55.

Miller, Merton H. "The Modigliani-Miller Propositions after Thirty Years." *Journal of Economic Perspectives*, Fall 1988, *2*(4), pp. 99–120.

______. *Merton Miller on derivatives*. New York: Wiley, 1997.

Modligiani, Franco. "Life Cycle, Individual Thrift, and the Wealth of Nations." *American Economic Review*, June 1986, *76*(3), pp. 297–313.

Parkinson, Michael. "Option Pricing: The American Put." *Journal of Business*, January 1977, *50*(1), pp. 21–36.

Richard, Scott F. "An Arbitrage Model of the Term Structure of Interest Rates." *Journal of Financial Economics*, March 1978, *6*(1), pp. 33–57.

Roll, Richard. *The behavior of interest rates: The application of the efficient market model to U.S. Treasury bills*. New York: Basic Books, 1970.

Ross, Stephen A. "Options and Efficiency." *Quarterly Journal of Economics*, February 1976, *90*(1), pp. 75–89.

Rubinstein, Mark E. "The Valuation of Uncertain Income Streams and the Pricing of Options." *Bell Journal of Economics*, Autumn 1976, *7*(2), pp. 407–25.

Samuelson, Paul A. "Rational Theory of Warrant Pricing." *Industrial Management Review*, Spring 1965a, *6*(2), pp. 13–31.

______. "Proof That Properly Anticipated Prices Fluctuate Randomly." *Industrial Management Review*, Spring 1965b, *6*(2), pp. 41–49.

Samuelson, Paul A. and Merton, Robert C. "A Complete Model of Warrant Pricing That Maximizes Utility." *Industrial Management Review*, Winter 1969, *10*(2), pp. 17–46.

Scholes, Myron S. "The Market for Securities: Substitution versus Price Pressure and the Effects of Information on Share Prices." *Journal of Business*, April 1972, *45*(2), pp. 179–211.

______. "Taxes and the Pricing of Options." *Journal of Finance*, May 1976, *31*(2), pp. 319–32.

______. "The Future of Futures," in William H. Beaver and George Parker, eds., *Risk management problems & solutions*, New York: McGraw-Hill, 1995, pp. 349–69.

______. "Global Financial Markets, Derivative Securities, and Systemic Risks." *Journal of Risk and Uncertainty*, May 1996, *12*(2–3), pp. 271–86.

Scholes, Myron S. and Wolfson, Mark A. *Taxes and business strategy: A planning approach*. Englewood Cliffs, NJ: Prentice Hall, 1992.

Sharpe, William F. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk." *Journal of Finance*, September 1964, *19*(3), pp. 425–42.

______. *Investments*. Englewood Cliffs, NJ: Prentice Hall, 1978.

Smith, Clifford W. and Stulz, René M. "The Determinants of a Firm's Hedging Policies." *Journal of Financial and Quantitative Analysis*, December 1985, *20*(4), pp. 391–405.

Stulz, René M. "Rethinking Risk Management." *Journal of Applied Corporate Finance*, Fall 1996, *9*(3), pp. 8–24.

Vasicek, Oldrich A. "An Equilibrium Characterization of the Term Structure." *Journal of Financial Economics*, November 1977, *5*(2), pp. 177–88.

Unraveling in Matching Markets

By HAO LI AND SHERWIN ROSEN*

We use a two-period matching model with initial uncertainty about productivities of participants to analyze incentives for early contracting or unraveling. Unraveling provides insurance in the absence of complete markets, but causes inefficient assignments. Unraveling is more likely the smaller the applicant pool, the smaller the proportion of more-promising applicants, and the greater the heterogeneity in the pool. Banning early contracts hurts firms and benefits less-promising applicants; the effects on more-promising applicants depend on how the gains from early contracts are shared. Ex post buyouts eliminate inefficient assignments, and more-promising applicants always unravel. (*JEL* D40, J44, C78)

The timing of transactions is important in markets where buyers and sellers need to be matched. There are many examples. Elite colleges have both regular- and early-admission programs. Most graduate- and undergraduate-admission procedures agree not to inform applicants prior to a common date. New entrants into many professional sports choose when to expose themselves to the draft. Much attention has been focused on the trend toward recruiting younger and less-experienced players in recent National Basketball Association (NBA) drafts. In 1997 the National Basketball Association draft set a record in that the first seven picks and 17 among 29 first-round selections were not college seniors. In the past few years some highly ranked players have skipped college altogether. In spite of the posturing by NBA executives urging players to stay in school and finish their education, 1997's outcome is along the creeping trend toward earlier entrance into the NBA.

Difficulties in controlling the timing of recruiting new entrants of highly trained professionals have been pointed out in a set of interesting papers by Alvin Roth (1984, 1991) and Susan Mongell and Roth (1991). Roth and Xiaolin Xing (1994) observe that these difficulties appear in many other markets, including the medieval weekly markets for ordinary commodities, medical interns and residents, and postseason college football bowl games.

Two related aspects of timing of market transactions must be distinguished. In one, labeled "jumping the gun" by Roth and Xing, participants compete for a limited supply of the best-qualified candidates or best positions in strategically timing their offers. For instance, firms sometimes make exploding offers that quickly expire, while job candidates often try to delay making a choice from available offers in the hope of receiving better ones. Another aspect of timing is unraveling of the appointment date. In some entry-level professional labor markets, employment begins only after attainment and certification of professional qualifications, but the appointment date sometimes unravels to a few years before that. Unraveling is defined as future employment contracts that are signed long before employment is to begin. Examples include the placement of medical interns and residents before the 1950's, summer job programs for law students prior to offers of longer-term positions upon graduation, and early commitment by judges to clerks. Often there are natural dates when the potential of candidates can be best assessed. Unraveling of the appointment date

---

* Li: School of Economics and Finance, University of Hong Kong, Pokfulam Road, Hong Kong (e-mail: haoli@econ.hku.hk); Rosen: Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637 (e-mail: s-rosen@uchicago.edu). The authors are grateful for the comments and suggestions made by Alvin Roth, Wing Suen, and anonymous referees.

is not so much about competing through strategically timing proposals and acceptances. It is more about *ex post* inefficiencies caused by making early contracts with incomplete information.

The two aspects are related. Sometimes jumping the gun can be the principal reason for unraveling of the appointment date. Still, isolating the insurance aspect of "futures" contracts from the strategic aspect helps us understand how market imperfections affect timing problems. It also allows the social costs and benefits of early contracting to be assessed. This paper analyzes price competition and market unraveling based on incomplete information about agents' future productive characteristics. Agents have to deal with both individual uncertainty about their traits, as well as aggregate uncertainty about the market value of those traits. Aggregate uncertainty arises in these markets because assignments are indivisible. If appointments are made after individual uncertainty is resolved, *ex post* efficiency of assignments is guaranteed. But agents are exposed to payoff risks because they can be on the long or short side of the market. Complete Arrow-Debreu securities markets would eliminate payoff risks to agents and guarantee *ex post* efficient job assignments. In the absence of such markets, unraveling (early contracting) can bring about limited insurance at the expense of *ex post* inefficient assignments. Unraveling is a manifestation of risk aversion under incomplete insurance markets. It relieves some of the anxiety about availability of jobs for applicants and of qualified candidates for firms.

We analyze markets where individual uncertainty applies only to job applicants, and consider two contractual situations—one where firms are bound by *ex post* unsuccessful early contracts, and another where buyouts of inefficient matches by firms are allowed. Without buyouts, unraveling need not occur. It is more likely the smaller the proportion of more-promising candidates in the applicant pool, and the smaller the total applicant pool relative to the number of positions. In our model, payoff risks to firms are the main source of insurance gains from early contracts. These gains are large when the applicant pool or proportion of promising applicants is small, because firms are less likely to fill their positions in the spot market. An increase in the degree of heterogeneity increases the likelihood that more-promising applicants unravel. In this model, unraveling always reduces the probability that productive applicants will be short and firms will be long in the *ex post* spot market. Therefore a ban on early contracts unambiguously decreases the *ex ante* welfare of firms. Since less-promising applicants never receive any rents from early contracts, the ban unambiguously improves their welfare. Whether the ban increases or decreases welfare of more-promising applicants depends on the gains they receive.

When buyouts of *ex post* unsuccessful early contracts are allowed, unraveling is more prevalent than when early contracts are binding. Buyouts amount to an up-front contingency clause in early contracts specifying the terms under which contracts are terminated *ex post*. Buyout provisions benefit participants as a whole because they eliminate *ex post* inefficient assignments and increase total gains from trade. Buyouts always increase the welfare of those less-promising candidates who would choose to wait when *ex ante* contracts are binding. Whether buyouts increase or decrease, the welfare of firms and more-promising candidates depends on how the gains they receive with buyouts compare with the gains they receive without them.

Section I introduces a two-period assignment model, defines individual and aggregate uncertainties that are central to our analysis of unraveling, and identifies the market imperfections that cause unraveling to occur. Section II analyzes unraveling in a competitive market equilibrium without buyouts, and considers how relative size, composition, and the degree of heterogeneity in the applicant pool affect unraveling. Section III considers unraveling when buyouts are allowed. The last section contains a brief summary of this paper and some comments on our findings.

I. The Assignment Market: Background

There are two models of matching: the assignment market (Tjalling Koopmans and Martin Beckmann, 1953; Lloyd Shapley and Martin Shubik, 1972), and the marriage prob-

ir posi-
in the
likeli-
unravel.
ces the
will be
ost spot
acts un-
lfare of
s never
the ban
velfare.
ses wel-
ends on

ful early
is more
re bind-
contin-
ying the
nated *ex*
ipants as
ost inef-
cal gains
the wel-
ates who
contracts
se or de-
d more-
how the
pare with

d assign-
aggregate
nalysis of
et imper-
cur. Sec-
mpetitive
, and con-
n, and the
icant pool
ers unrav-
ie last sec-
paper and

ground

ng: the as-
omans and
hapley and
riage prob-

---

lem (David Gale and Shapley, 1962). These exchanges differ from others in that objects of trade are indivisible. Typically, each participant has only one unit to buy or sell. In an assignment market, money is transferable among all participants and matching is made through market prices. In a marriage problem, only rank orderings of preferences matter and prices play no allocating role.[1] We use the assignment market and competitive mechanism as the framework of our analysis.

A. *Competitive Equilibrium in the Assignment Market*

In an assignment market, there are two groups of agents: workers and firms. A production function describes a nonnegative joint output of an assignment of a worker to a firm. The joint output of assigning either two workers or two firms to each other is zero. The output of an unassigned agent is normalized to zero. The largest total output of any number of agents is the maximal sum of outputs that can be produced by pairwise assignments of these agents.

A pairwise assignment is feasible if each agent is matched to at most one other agent. Associated with a feasible assignment is a feasible outcome, a vector of nonnegative payoffs such that the sum of the payoffs to each matched pair is no greater than the joint output of the pair, and the payoff to each unmatched agent is zero. An assignment is efficient if among all feasible assignments it yields the largest sum of joint outputs. A feasible outcome is stable if the sum of the payoffs to any pair of agents is greater than or equal to the joint output they can produce. Thus, in a stable outcome, the sum of payoffs equals the joint output for each matched pair. For any pair not matched to each other, the sum of payoffs for the two agents is greater than the joint output they can produce, so that the pair does not form a "blocking" coalition that improves the payoff for both of them.

The main theorem of the assignment market is that if a feasible outcome is stable then the associated assignment is efficient, and if a feasible assignment is efficient then there exists a feasible outcome associated with it that is stable. This is analogous to the two welfare theorems in a decentralized economy. Koopmans and Beckmann (1953) formulate the efficient assignment as the solution to a linear programming problem and give a competitive market equilibrium interpretation to the associated outcome. Shapley and Shubik (1972) formulate stable outcomes as equilibria of a cooperative game and show that side-payments do not occur in equilibrium. In this paper, we refer to each stable outcome as a competitive equilibrium.

B. *Payoff Risks*

Parallels between the assignment theory and competitive markets are best revealed by an example with two distinct types of identical agents in each group. Suppose that workers and firms are either productive or unproductive, and that assigning a productive worker to a productive firm produces unit output, while all other kinds of assignments produce nothing.

To put the assignment market in a competitive framework, it is useful to think of one group of agents, say the firms, as residual income recipients. Unproductive firms cannot bid positive prices for workers and receive a payoff of zero in the market. Similarly, unproductive workers cannot ask positive prices of firms and receive zero payoff. Only productive workers and productive firms potentially have positive payoffs. Their actual payoffs in a competitive market depend on their relative numbers. If there are equal numbers of productive workers and firms, all are matched one-to-one. Any division of the pairwise output that gives all agents in each group the same payoff is a competitive equilibrium. Suppose, however, that there are more productive workers than firms. Now the supply of productive workers exceeds demand and their payoff is driven down to zero. Productive firms get all the return because they are the scarce factor. In the opposite case where there are more productive firms than workers, the workers receive 1 in the competitive equilibrium. In our simple example of two types o each group, uncertainty about the relative size of the two productive types translates into payoff risks before they enter the market. Avoiding these kind of payoff risks is a main motivation for individual agents to contract early and "unravel" the market.[2]

---

[1] For a discussion of modeling issues in the two problems, see Vincent Crawford, 1991.

Discontinuities in equilibrium prices or payoffs with respect to supply and demand are inherent in the indivisibility of assignments (pairwise here) and in the discreteness in the types. Indivisibility implies that payoffs are sensitive only to the sign of excess demand or supply, not to its magnitude. Whether productive workers outnumber productive firms by a few or by many does not change the equilibrium payoffs of zero for productive workers and 1 for productive firms. The discontinuities do not disappear as the size of the market gets large. They are more sensitive to the discreteness of types. For example, suppose the joint output between an unproductive agent and a productive agent of the other group is $\mu \in (0, 1/2)$, instead of 0. Then the consequences of being on the long or short side of the market are less severe. If productive workers outnumber productive firms but there are more firms than productive workers, so that each productive worker is matched to some firm, their equilibrium payoff is μ instead of zero because they add value μ to assignments with unproductive firms. Productive firms receive $1 - \mu$ instead of 1. In the limit of a continuum of types, the discontinuities disappear.

C. *An Example of Unraveling*

The essential aspects of market unraveling are illustrated by a two-period model with both individual uncertainty and aggregate uncertainty. Individual uncertainty arises because there is incomplete information in the first period by all agents on how productive a particular agent will be in the second period. Individual uncertainty creates randomness in supplies of different types of agents, and results in aggregate uncertainty due to discontinuities in equilibrium payoffs. Unraveling (first-period contracting) eliminates aggregate uncertainty but causes *ex post* inefficient assignments due to individual uncertainty. Contracting after individual uncertainty is resolved ensures *ex post* efficient assignments, but exposes agents to payoff risks due to aggregate uncertainty.

These points are illustrated by extending the example. In the first period, productivities of agents are unknown. All agents on each side of the market are identical with probability λ of becoming productive in the second period. All are risk averse with utility function u. For simplicity, we assume that there are equal numbers of workers and firms.

The second-period competitive equilibrium gives agents who turn out to be productive a payoff of 1 or 0 depending on excess demand or supply.[3] The uncertainty of second-period payoffs may motivate agents to contract with each other in the first period, before their productivity is known. Imagine the following game. In the first period, there are pairwise meetings between workers and firms. Each pair of agents chooses whether or not to sign the following contract: the worker receives a payoff r and the firm receives $1 - r$ in the second period if both turn out to be productive; otherwise, both receive 0. The market is said to unravel if there exists some $r \in [0, 1]$ such that some pairs of agents choose to sign the first-period contract rather than to wait. From the individual point of view, the first-period contract eliminates risks that the two agents may be on the long or short side of the market, should they be productive. But the contracts create *ex post* assignment inefficiencies because information necessary to achieve efficient assignments is available only in the second period after the contracts have been signed. The cost of these inefficiencies to the two agents is reflected in the fact that the payoff to the two agents depends on the joint probability that both are productive.

---

[2] Roth and Xing (1994) identify two other causes of unraveling. One is instability: if the assignment in the market is unstable, then there may be incentives for agents to reach agreements early. However, this cannot happen in a competitive assignment market. They identify market competition as another cause of unraveling and give an example where despite the stability of assignment in the market, some agents have incentives to make early offers to their next-best candidates by committing to not competing for these agents in the market in case they reject the offers.

[3] The equilibrium payoffs are indeterminate if there are equal numbers of productive agents. In this case, we assume that the equilibrium payoff to all productive workers is 0 or 1 with probability $^1/_2$. This assumption is nonconsequential for what follows.

When λ is sufficiently large the insurance gains from contracting early and eliminating aggregate uncertainty outweigh the loss due to *ex post* inefficient assignments; the market unravels. The argument runs as follows. In the second period, the payoffs to productive agents are 1 or 0 depending on excess demand or supply. The expected utility to each agent from waiting is $\lambda[u(1)/2 + u(0)/2] + (1 - \lambda)u(0)$. If two agents sign the first-period contract the expected utility to the worker is $\lambda^2 u(r) + (1 - \lambda^2)u(0)$ and the expected utility to the firm is $\lambda^2 u(1 - r) + (1 - \lambda^2)u(0)$. Let r^w be the minimum payoff required for the worker to sign the contract and r^f be the maximum payoff that the firm is willing to pay the worker to sign the contract. Equating the expressions above we have that the ask price r^w and the bid price r^f are implicitly defined by the condition

$$(1) \quad u(r^w) = u(1 - r^f) = u(0) + [u(1) - u(0)]/(2\lambda).$$

If λ is large enough then the ask price r^w falls below the bid price r^f and unraveling occurs. There is not enough structure in this example to determine the market equilibrium price uniquely, but it must be between the bid and ask. When λ is sufficiently small, $r^f < r^w$; the market does not unravel. The value of insurance from a first-period contract is small because the probability of exposure to aggregate uncertainty in the second period is small. The loss due to *ex post* inefficient assignments dominates the insurance gains.

D. *Unraveling and Market Incompleteness*

Before we analyze richer assignment problems, it is important to understand that unraveling is really caused by market incompleteness. In the above example, the optimal insurance arrangement is for all agents to sign a first-period contract in which they all receive the same share of maximum total output of all participants after waiting until the second period to achieve *ex post* efficient assignments. A decentralized Arrow-Debreu securities market in the first period together with a second-period spot market achieves this.

Imagine that in the first period each agent sells 100 perfectly divisible shares of the claim to the agent's payoff from the second-period spot market. For example, holding 50 shares of one claim entitles the buyer to one-half of the seller's equilibrium payoff from the spot market. A competitive equilibrium in the securities market is the set of share prices, such that given these prices, each agent chooses the numbers of shares to maximize expected utility subject to the budget constraint, and the market clears for shares of each claim. Since all agents are indistinguishable in the first period, the competitive equilibrium has equal share prices for all claims and each agent holds an equal number of shares of each claim, including the agent's own claim. This equilibrium is *ex post* efficient, because the second-period spot market produces the efficient assignments that maximize total output. It is also *ex ante* optimal, because each agent spreads risks as much as possible.

If markets were complete and agents could legally sell claims to their future payoffs, unraveling would not be observed. Unraveling or early contracting is a manifestation of market failure. For instance, it is never said to be observed in markets for agricultural commodities where futures contracts are common. Rather, it is restricted to those labor markets where indivisible assignments are important and where complete markets of the kind described above are vulnerable to moral hazard problems. These problems are especially severe in labor markets because *ex post* productivities of job applicants depend critically on the efforts they make after signing early contracts. Insurance provided by complete markets may lead to shirking and disinvestment in human capital by the applicants. In what follows we take it for granted that complete markets do not exist, so that unraveling may occur.

II. Unraveling Without Reentry

In this section, we consider a richer version of the example above. There agents are indistinguishable in the first period and unraveling affects all market participants in the same way. Here we allow for two kinds of heterogeneity among participants. First, the two sides in the assignment market differ in the individual uncertainty they face. In markets for entry-level professionals, individual uncertainty about future productivity is a substantial issue only for job applicants: Firms operate each year and have established reputations. Second, there is heterogeneity among the workers. Some have more promise to become productive later than others. Unraveling in this situation can affect more-promising workers differently than less-promising ones. The competitive equilibrium in the first-period market is characterized next, assuming that agents who sign first-period contracts do not enter the spot market in the second period. The case of unraveling with buyouts and reentry is considered later.

A. *The Model*

As before, in the second period workers are either productive or unproductive, and the joint output is 1 between a productive worker and a productive firm and 0 otherwise. In the first period, there are n types of workers, whose prospects differ. Type i ($i = 1, \ldots, n$) workers, of size n_i, become productive in the second period with probability λ_i. Type 1 workers are the most promising. Less-promising types (greater i) have smaller λ_i. To capture the idea that firms face little uncertainty about themselves, we assume that all firms are identical and all are productive. All workers have the same concave utility function u over payoffs, and all firms have the same utility function v. Without loss of generality, assume that $u(0) = v(0) = 0$.

The equilibrium offers written in the first-period contracts and the number of each type of agents who sign them are endogenous. For agents who choose to wait, the second-period spot market is described as before: productive agents receive a payoff of 1 if they are on the short side of the market. All other agents receive zero. Since there is no reentry, the first-period contract between a type i worker and a firm takes the following form: any type i worker who turns out to be productive receives r_i and the firm receives $1 - r_i$. Otherwise both receive zero. We refer to r_i as the "price" of the first-period contract with type i workers.[4]

We proceed by defining bid and ask prices for the workers. Different types of workers may have different bid and ask prices. These prices depend on the number y_i of type i workers who sign first-period contracts because aggregate uncertainty—the discontinuous payoffs in the second-period spot market conditional on excess demand or supply—depends on y_i. For any $y_1, \ldots, y_n$, let $\pi(y_1, \ldots, y_n)$ be the probability that there are more firms in the second-period spot market than productive workers. That is, $\pi(y_1, \ldots, y_n)$ is the probability that *ex post* productive workers who have not signed early contracts receive 1 in the second-period spot market;[5] it represents aggregate uncertainty. To shorten the notation, we write π instead of $\pi(y_1, \ldots, y_n)$ whenever the meaning is clear.

B. *Bid and Ask Prices*

Given $\pi \in [0, 1]$, type i workers are indifferent between signing a first-period contract at price r_i and waiting for the second-period spot market, if

$$(2)\quad (1 - \lambda_i)u(0) + \lambda_i u(r_i) = (1 - \lambda_i)u(0) + \lambda_i[\pi u(1) + (1 - \pi)u(0)].$$

This simplifies to

$$(3)\qquad u(r_i) = \pi u(1).$$

---

[4] In general an optimal first-period insurance contract requires nonzero transfer between the two parties when the worker turns out to be unproductive. Our assumption that firms do not pay *ex post* unproductive workers can be justified on the ground that the worker's *ex post* productivity depends on some effort he makes between the two periods. It also makes it easy to characterize the first-period contracts. Subsection E relaxes this assumption.

[5] For simplicity, indeterminacy of the spot market equilibrium is resolved by our assumption that when there are equal numbers of productive workers and firms, the equilibrium is the same as when there are fewer productive workers than firms.

Therefore, the ask price of all types of workers is the same. Let $r^w(\pi)$ denote the common ask price as a function of π. In order for firms to be indifferent between signing a first-period contract with a type i worker at price r_i and waiting for the second-period spot market, we must have

$$\lambda_i v(1 - r_i) = (1 - \pi)v(1). \tag{4}$$

Let $r_i^f(\pi)$ denote the bid price for type i workers. For any types $i < j$, $r_i^f(\pi) \geq r_j^f(\pi)$ for all π, with equality if and only if $\pi = 1$.

Taking derivatives with respect to π, we have

$$\frac{dr^w(\pi)}{d\pi} = \frac{u(1)}{u'(r^w(\pi))}; \tag{5}$$

$$\frac{dr_i^f(\pi)}{d\pi} = \frac{v(1)}{\lambda_i v'(1 - r_i^f(\pi))}. \tag{6}$$

Thus, $r^w(\pi)$ is increasing and convex and $r_i^f(\pi)$ is increasing and concave. One can easily verify that $r^w(0) = 0$, $r_i^f(0) < 0$, and $r^w(1) = r_i^f(1) = 1$ for each i.

Figure 1 depicts the relative positions of the ask price r^w and bid prices r_i^f and r_j^f for $i < j$. In the diagram, r^w intersects r_i^f at $\pi_i \in (0, 1)$ and it intersects r_j^f at $\pi_j \in (\pi_i, 1)$. $r^w(\pi) > r_i^f(\pi) > r_j^f(\pi)$ for $\pi \in [0, \pi_i)$, $r_i^f(\pi) > r^w(\pi) > r_j^f(\pi)$ for $\pi \in (\pi_i, \pi_j)$, and $r_i^f(\pi) > r_j^f(\pi) > r^w(\pi)$ for $\pi \in (\pi_j, 1)$. Therefore, when $\pi > \pi_j$, insurance gains exist from first-period contracts with both type i and type j workers. Between π_i and π_j insurance gains exist for type i workers but not for type j. Below π_i the ask price exceeds both bid prices so there is no gain from insurance for either type. That insurance gains exist only when π is high is due to the absence of individual uncertainty for firms. When π is very low, the ask price of both types of workers is close to zero, but because of the individual uncertainty of workers, the bid price for each type is negative: firms prefer to wait for the spot market rather than to sign first-period contracts with either type of worker at any price. In this sense we can say that payoff risks to firms are the main source of insurance gains in our model.

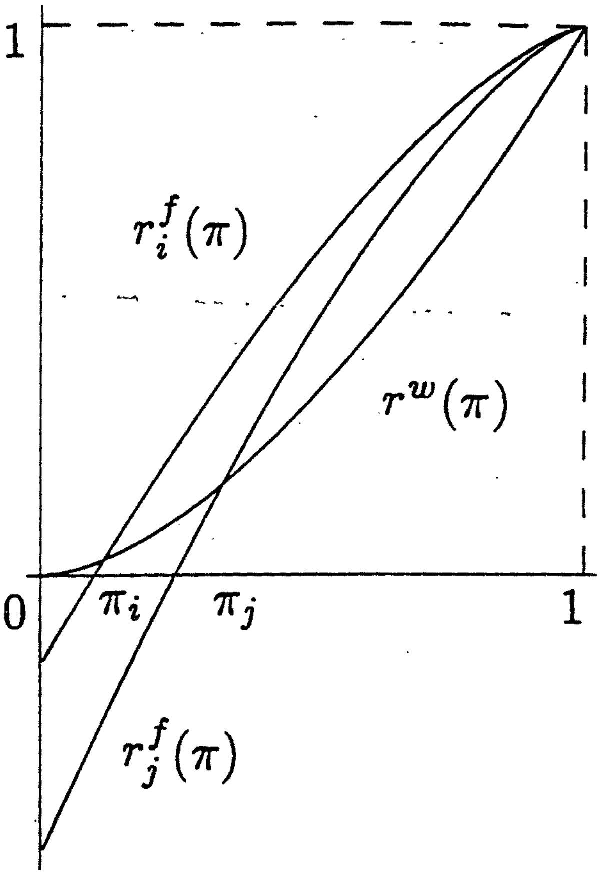


FIGURE 1. ASK AND BID PRICE FUNCTIONS WITHOUT REENTRY

For a given type i, the ask price function $r^w(\pi)$ may lie entirely above the bid price $r_i^f(\pi)$ in Figure 1 for all $\pi < 1$. Then type i workers never participate in unraveling. In this case, we define the threshold π_i to be 1. Whether π_i is equal to 1 or strictly less than 1 depends on the curvatures of r^w and r_i^f, and on the probability λ_i if u or v is concave. The curvatures of r^w and r^f in turn depend on risk aversion. If both u and v are linear (risk neutral), $dr^w/d\pi = 1$ and $dr_i^f/d\pi = 1/\lambda_i$: there are no gains from first-period contracts regardless of λ_i. The more risk averse the agents, the more likely gains from insurance. If u or v are concave, the bid price for type i workers is increasing in λ_i, and insurance gains are more likely to exist if λ_i is greater.

C. *Ordering Property and Monotonicity Property*

It is useful to think of the first-period market as an assignment market. If a worker and firm sign a first-period contract, they are said to be assigned to each other. If a worker or a firm chooses to wait for the second-period market, the agent is said to be unassigned. In an assignment problem without uncertainty, the payoff to an unassigned worker is exogenous. Here it depends both on the type of the worker and on y_i, because these numbers affect aggregate uncertainty in the second-period spot market. Furthermore, the numbers are endogenous variables, determined in the first-period market equilibrium.

Unraveling is signing a first-period contract. Type i workers unravel if the bid price exceeds the ask price. The market equilibrium price r_i lies between the bid and ask. Unraveling of type i workers is complete if all type i workers sign first-period contracts: $y_i = n_i$. Unraveling of type i is incomplete if $y_i < n_i$. In this case all type i workers are indifferent between signing and waiting for the second-period spot market. If firms sign first-period contracts with more than one type of worker, contract prices must be such that firms are indifferent between all these types. If some firms sign while others wait, we say there is incomplete unraveling of firms, and firms are indifferent between signing and waiting. We assume that the total number of n types of workers exceeds the number of firms.[6]

Two properties of the model greatly simplify the characterization of market equilibrium. First, unraveling is ordered: unraveling of type i workers "precedes" that of type $j > i$ workers. This ordering property limits the number of corner solutions that must be considered. Second, the probability $\pi(y_1, \dots, y_n)$ is monotonically decreasing in each of its arguments. This monotonicity property further simplifies characterization of equilibrium.

Consider two types of workers, i and $j > i$. The ask price for the two types is the same. Since type i has a greater probability of becoming productive than type j, the bid price for i is greater than that for j for any $\pi < 1$. Thus, unraveling of type j implies complete unraveling of type i. The reason is as follows. Since there is unraveling of type j, the price r_j paid by firms for contracts with type j is higher than the ask price $r^w(\pi)$. If there was incomplete unraveling for type i, the price r_i for the contract with type i would equal the ask price $r^w(\pi)$, implying $r_i \leq r_j$. But then $\lambda_i v(1 - r_i) > \lambda_j v(1 - r_j)$ because $\lambda_i > \lambda_j$, and firms strictly prefer signing first-period contracts with type i to signing with type j, contradicting the requirement that firms be indifferent between two types of workers if they sign with both.

The absence of individual uncertainty for firms implies that $\pi(y_1, \dots, y_n)$ is a decreasing function in each of its arguments. If an additional pair of a type i worker and a firm sign a first-period contract, the number of firms remaining in the second-period spot market falls by one. Individual uncertainty for type i workers means that the expected number of productive workers participating in the second-period market falls by less than one. Therefore, the probability that productive workers will be short in the second-period spot market decreases with y_i, the number of type i workers who sign first-period contracts. The formal argument is straightforward and omitted.

D. *Market Equilibria*

Comparisons between aggregate uncertainty π and the thresholds π_i determine the unraveling equilibrium in the first-period market. Recall from Figure 1 that π_i is defined as the intersection between the bid and ask functions of type i workers, and that less-promising types have a greater π_i. Gains from insurance for type i exist only if π exceeds π_i. Aggregate uncertainty depends on the number y_i of each type i workers who sign first-period contracts and on the size n_i of type i. The ordering property implies that in any equilibrium there is a cutoff, least-promising type i_* of workers that participate in unraveling. Recall that $\pi(0, \dots,$

---

[6] If there are more firms than workers of all types, there is no insurance gain from early contracts because $\pi = 1$ and workers will definitely be short in the spot market. (Insurance gains still exist if firms offer full insurance to workers; see subsection E.) However, it is misleading to say that excess supply of workers is critical for unraveling to occur. What is more important in the early contracting market is the expected supply of productive workers. Indeed, Figure 1 shows that expected shortage of productive workers in the sense of relatively big π is necessary for unraveling.

0) is the probability that productive workers are short in the second-period spot market when no one signs a first-period contract. To find i_*, we first compare $\pi(0, \dots, 0)$ to π_1. If $\pi(0, \dots, 0) \leq \pi_1$, set $i_* = 0$. Otherwise, compare $\pi(n_1, 0, \dots, 0)$ to π_2. If $\pi(n_1, 0, \dots, 0) \leq \pi_2$, set $i_* = 1$. Otherwise, compare $\pi(n_1, n_2, 0, \dots, 0)$ to π_3. If $\pi(n_1, n_2, 0, \dots, 0) \leq \pi_3$, set $i_* = 2$. Otherwise, compare $\pi(n_1, n_2, n_3, 0, \dots, 0)$ to π_4, and so on. This process stops before we reach the $(n + 1)$-th step, because the total number of workers exceeds the number of firms.

If the process stops in the first step and $i_* = 0$, there is no unraveling of any type. At and below the initial condition $\pi(0, \dots, 0) \leq \pi_1$, there is no insurance gain for type 1, hence no insurance gain for any other types. Unraveling of type 1 would only cause the aggregate uncertainty π to decrease further. If $i_* \geq 1$, by construction it satisfies

$$(7) \quad \pi(n_1, \dots, n_{i_*-1}, 0, \dots, 0) > \pi_{i_*};$$

$$(8) \quad \pi(n_1, \dots, n_{i_*}, 0, \dots, 0) \leq \pi_{i_*+1}.$$

The first inequality implies positive insurance gain for type i_* workers after all more-promising types have unraveled; the second shows that insurance gains for type $i_* + 1$ do not exist after type i_* and all the more-promising types have unraveled. By the ordering property, type i_* is the least-promising type of workers who unravel.

The equilibrium prices and π depend on whether unraveling of the cutoff type i_* workers is complete or incomplete. If $\pi(n_1, \dots, n_{i_*}, 0, \dots, 0) \geq \pi_{i_*}$, unraveling of type i_* is complete. Insurance gains for type i_* workers still exist at the point where all of them have signed early contracts, but no more of them are available. The equilibrium π is $\pi(n_1, \dots, n_{i_*}, 0, \dots, 0)$. There is incomplete unraveling of firms, which are indifferent between signing first-period contracts with types $i = 1, \dots, i_*$ and waiting for the second-period spot market.[7] The equilibrium price of the first-period contract for type i workers ($i = 1, \dots, i_*$) is $r_i = r_i^f(\pi(n_1, \dots, n_{i_*}, 0, \dots, 0))$, the bid price for type i evaluated at the equilibrium π. Workers who unravel receive all the rent from first-period contracts.

If $\pi(n_1, \dots, n_{i_*}, 0, \dots, 0) < \pi_{i_*}$, unraveling of type i_* is incomplete. Unraveling continues until the equilibrium π is driven down to π_{i_*}.[8] All the insurance gains from first-period contracting are exhausted at this point. Again, the equilibrium price r_i of the first-period contract for type i workers ($i = 1, \dots, i_*$) is given by $r_i = r_i^f(\pi_{i_*})$, on the bid function for type i, and workers more promising than type i_* receive all the rent. Both type i_* workers and firms are indifferent between signing and waiting, but more-promising types of workers strictly prefer signing to waiting.

E. *Unraveling with Full Insurance*

In the first-period contracts considered above, payments received by applicants depend on their *ex post* productivity. This kind of contract has the advantage of preserving workers' incentives when *ex post* productivity of applicants depends critically on the efforts they make after signing early contracts. Firms do not fully insure applicants in this case, even if they are risk neutral. Nonetheless, it is interesting to analyze first-period contracts that provide full insurance. For example, if investments by applicants after first-period contracts are not firm specific, then the applicants pursue their own long-term interests by making the investment. In these situations, risk-neutral firms may be willing to offer full insurance to applicants in early contracts.

When risk-neutral firms offer full insurance to workers, the first-period contract with a type i ($i = 1, \dots, n$) worker specifies a transfer payment r_i to the worker regardless of his *ex post* productivity. For any given aggregate

---

[7] Unraveling for firms cannot be complete because there are no insurance gains when aggregate uncertainty π equals 0.

[8] We ignore the integer problem and assume that y_i's can take real values. More rigorously we can assume agents use mixed strategies in the sense that the probability of signing a first-period contract can be any number between 0 and 1.

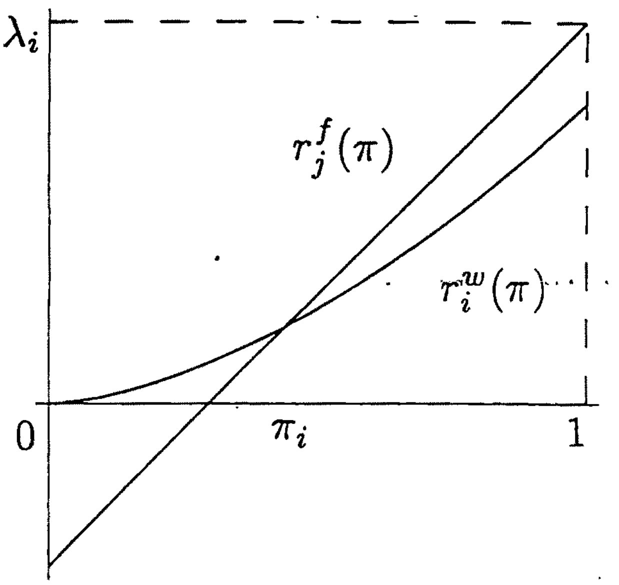


FIGURE 2. ASK AND BID PRICE FUNCTIONS WITH FULL INSURANCE

uncertainty $\pi \in [0, 1]$, the ask price $r_i^w(\pi)$ of type i is the solution to

$$(9) \qquad u(r) = \lambda_i \pi u(1).$$

The bid price for type i is given by

$$(10) \qquad r_i^f(\pi) = \lambda_i - (1 - \pi).$$

Ask and bid functions have the following properties: (i) both are increasing in π; (ii) ask price is zero and bid price is negative at $\pi = 0$, and bid price exceeds ask price at $\pi = 1$; and (iii) ask and bid functions have a unique intersection π_i.[9] [See Figure 2.]

The ask and bid functions can be compared to their counterparts in Figure 1 when firms are risk neutral (v is linear) but first-period contracts do not provide full insurance. We observe that there is more to gain when firms offer full insurance, even though both ask and bid functions shift down (with full insurance, workers ask less and firms bid less). Without full insurance, gains from insurance may not exist for any π if λ_i is too small (ask exceeds bid for all $\pi < 1$), but insurance gains always exist with full insurance for any $\lambda_i < 1$ if π is great enough. [See Figure 2.] Moreover, for any λ_i, the critical aggregate uncertainty π_i, below which insurance gains do not exist, can be shown to be lower with full insurance than without full insurance.

As before, insurance gains are greater for more-promising types: the critical aggregate uncertainty π_i is smaller for more-promising types. Therefore, the ordering property is preserved under full insurance. Unraveling of any type i implies complete unraveling of all types more promising than i. To see this, note that unraveling for firms cannot be complete because there are no insurance gains for any type when aggregate uncertainty π equals 0. Thus, in equilibrium firms are indifferent between all types of workers with whom they sign early contracts and waiting for the second-period market. This implies that equilibrium first-period contract prices for these types are given by their bid prices $r_i^f(\pi)$ evaluated at the equilibrium aggregate uncertainty. If some type j workers unravel, then $r_j^f(\pi) \geq r_j^w(\pi)$, implying $\pi \geq \pi_j$. Since $\pi_i < \pi_j$ for any type i more promising than type j, $\pi > \pi_i$, and so $r_i^f(\pi) > r_j^w(\pi)$. Therefore, unraveling of type i is complete.

Since the monotonicity property of aggregate uncertainty holds as before, we can find the equilibrium π, the cutoff type, and equilibrium prices in exactly the same way as in the previous section. The discussions about comparative statics analyzed next apply to both early contracts without full insurance and with full insurance.

F. *Comparative Statics*

Using the characterization of unraveling equilibrium established above, we now address how changes in relative numbers of different kinds of agents, the average quality of workers, and the degree of heterogeneity affect the likelihood of unraveling of different types, equilibrium prices, and *ex ante* welfare in terms of first-period expected utility.

From the ordering and monotonicity properties, the necessary and sufficient condition for unraveling of type i is $\pi(n_1, \dots, n_{i-1}, 0, \dots, 0) > \pi_i$. In an assignment market without uncertainty, equilibrium payoffs are determined by relative numbers of different types of market participants up to a boundary condition that those remaining unassigned receive an exogenous payoff. Here, the "effective" numbers of market participants $\pi(n_1, \ldots, n_{i-1}, 0, \ldots, 0)$ (relative to π_i) replace the sizes n_i as the determinants of equilibrium. The thresholds π_i do not depend on relative numbers of market participants, but $\pi(n_1, \ldots, n_{i-1}, 0, \ldots, 0)$ decreases as the number n_i of any type i workers increases or as the number of firms decreases. Therefore, unraveling of any type i workers becomes less likely as the relative supply of the n types of workers increases. Furthermore, given the total number of workers, unraveling of more-promising types is less likely the greater the proportion of more-promising types of workers in the distribution of types, because a greater proportion of more-promising types on balance decreases $\pi(n_1, \ldots, n_{i-1}, 0, \ldots, 0)$ for any small i.

---

[9] We use the same notation to ease the burden.

The effects of numbers of market participants on equilibrium prices and on *ex ante* welfare of market participants depend on whether unraveling of the cutoff type i_* is complete or not. Relative numbers do not affect bid and ask functions, so if unraveling of type i_* is incomplete, changes in the numbers of different types of market participants on the margin have no effects on prices and hence no effects on the expected utility of any type of agents who enter early contracts. However, if unraveling of type i_* is complete, changes in the numbers of market participants can affect the equilibrium prices through $\pi(n_1, \ldots, n_{i_*}, 0, \ldots, 0)$. Increasing the number of any type i of workers and decreasing the number of firms have the same effect of reducing the equilibrium prices r_i for types $i \leq i_*$ because they reduce $\pi(n_1, \ldots, n_{i_*}, 0, \ldots, 0)$. The *ex ante* welfare of all types of workers who unravel is decreased while welfare of firms is increased. Types that in equilibrium do not unravel are also worse off, because the equilibrium π, the probability of *ex post* productive workers receiving 1 in the second-period spot market, falls.

Across-the-board changes in the prospects λ_i of applicants have ambiguous effects on the likelihood of unraveling. With an increase in all λ_i, the bid function for each type i shifts up while the ask function stays the same, so π_i decreases and insurance gains increase for type i workers. [See Figure 1.] This tends to increase the likelihood of unraveling for type i workers. But an improvement in the prospects of all applicants also leads to a decrease in the probabilities $\pi(n_1, \ldots, n_{i-1}, 0, \ldots, 0)$ for all i. This tends to decrease the likelihood of unraveling for all types. Thus, changes in the likelihood of unraveling are ambiguous.

Similarly, changes in the degree of heterogeneity among workers affect both the average quality of workers available in the second-period spot market through changes in $\pi(n_1, \ldots, n_{i-1}, 0, \ldots, 0)$, and the insurance gains for individual types through changes in π_i. A "compensated" change in heterogeneity is one that isolates the two effects. For example, suppose initially $\pi(0, \ldots, 0) < \pi_1$ so that there is no unraveling. Consider an increase in λ_1 and a decrease in λ_n that increases heterogeneity among workers but keeps $\pi(0, \ldots, 0)$ unchanged. This decreases π_1 and increases the insurance gains for type 1. If the changes are sufficiently large, type 1 unravels. Moreover, $\pi(n_1, \ldots, n_{i-1}, 0, \ldots, 0)$ increases for all i. Therefore, the likelihood of unraveling for all types (except perhaps for type n) increases as a result of the increase in heterogeneity.

We conclude with the welfare effects of banning first-period contracting. Workers who do not unravel in equilibrium benefit from such a ban, because under the ban the probability that *ex post* productive workers receive 1 equals $\pi(0, \ldots, 0)$, which is greater than the equilibrium probability $\pi(n_1, \ldots, n_{i_*}, 0, \ldots, 0)$ or π_{i_*}. For the same reason, the ban hurts firms unambiguously. The welfare effects of the ban on workers who unravel can be ambiguous. These workers receive rents from signing first-period contracts rather than waiting. Whether they are made better off or worse off by the ban depends on the rent they receive and the impact the ban has on the difference between the equilibrium value of π and $\pi(0, \ldots, 0)$. More-promising types are more likely to be hurt by the ban because they receive bigger rents.

III. Unraveling with Reentry

The contracts analyzed above are inefficient in both the first period and the second period.

Pairwise first-period contracts cannot replicate the gains from risk sharing implied by complete contingent contracts. In the second period, *ex post* inefficiencies are caused by the restriction on recontracting for those *ex post* productive agents whose first-period contracts proved unsuccessful because their partners turned out to be unproductive. In this section, *ex post* productive agents are allowed to enter the second-period spot market. This amounts to allowing firms to "buy out" of their *ex post* unsuccessful first-period contracts.[10] Inefficient risk sharing remains in the first period, but *ex post* inefficiencies are eliminated because all *ex post* unsuccessful first-period assignments are dissolved.

A. *Buyout Provisions*

We need to analyze the terms of buyouts. Consider what happens when a type i worker ($i = 1, \dots, n$) in a first-period contract with a firm turns out to be productive. If the spot market happens to be long on productive workers and they receive a spot payoff zero, the firm must pay the first-period contract price of r_i to the type i worker anyway, so there is no advantage to buying out. Similarly, if the spot price for productive workers is 1, there is no advantage to buying out for the worker. Thus, the terms set by the initial contract cannot be advantageously renegotiated if the early contract turns out to be successful. Next, consider what happens when a first-period contract is successful. If the spot price for productive workers is 1, the firm and the worker gain nothing from buying out. Competition among firms implies that some of them will buy out of their contracts paying their *ex post* unproductive partners 0, and rematch to the *ex post* productive workers who have not signed first-period contracts paying them 1. But for the worker and the firm who have signed a first-period contract, it is as if the buyout had not taken place. Therefore, for these agents, meaningful buyouts occur only when the contract is unsuccessful and the spot price of productive workers is 0.

Because agents participating in a first-period contract anticipate the possibility of buyout, renegotiation amounts to an up-front contingency clause that specifies the terms and the market circumstances of a buyout. The firm pays the *ex post* unproductive type i worker r' to buy out of the contract when productive workers are available in the second-period spot market at the price of 0. Since the joint gain from the buyout is 1, the two agents in the first-period contract are in the same position as they would have been had the worker been productive. An optimal first-period contract should share risks in the same way. Therefore, $r' = r_i$.[11]

B. *Market Equilibria*

The analysis of equilibrium unraveling with reentry proceeds as before. In fact, the analysis becomes easier. With buyouts, all *ex post* inefficient assignments are eliminated and the key probability π that productive workers are short in the second-period market is independent of the number of workers y_i who sign first-period contracts. Thus π is no longer endogenous.

As in the previous section, we first characterize the bid and ask functions. For any fixed $\pi \in [0, 1]$ and $i = 1, \dots, n$, let $r_i^w(\pi)$ be the solution to the equation

$$(11) \quad [1 - (1 - \lambda_i)\pi]u(r) = \lambda_i \pi u(1),$$

---

[10] The reason why firms buy out unsuccessful contracts here is that they have no individual uncertainty: *ex post* productive workers who sign first-period contracts are always efficiently assigned to productive firms. Buyout will be two-sided if there is individual uncertainty for firms as well as for workers. See Li and Wing Suen (1997). Also, note that when firms are risk neutral and offer full insurance to workers, first-period contracts are necessarily binding because there is no room for renegotiation.

[11] When the cost of reentry is a number $k \in (0, 1)$ instead of 0 in this section and 1 in the previous section, equilibrium first-period contracts specify the division of the output 1 when the type i worker turns out to be productive and of the output $1 - k$ when he is not productive but productive workers are on the long side in the second-period market. The payoff risks in these two contingencies must be optimally shared between the two parties. The analysis of unraveling with costly reentry is similar to that in this section. The only change is that insurance gains may not exist when π is small because reentry is likely and the cost of reentry reduces the value of first-period insurance contracts.

and let $r_i^f(\pi)$ be the solution to the equation

$$(12)\qquad [1-(1-\lambda_i)\pi]v(1-r) = (1-\pi)v(1).$$

With reentry, each type of workers has distinct ask and bid functions. It is straightforward to establish that for each type i of worker: (i) ask and bid price functions are increasing in π; (ii) ask and bid prices equal zero at $\pi = 0$ and equal 1 at $\pi = 1$; and (iii) bid price is strictly greater than ask price for all π between 0 and 1. [See Figure 3.] Unlike the previous section, insurance gains exist for all values of π for all types of workers.

Without reentry, ask prices are the same for all types of workers and bid price is greater for more-promising types. The value of insurance is greater for more-promising types and unraveling of workers is ordered by type. Here, both the bid and ask prices are greater for type i workers than for type $j > i$ at each value of π. It can be shown that although type j workers are willing to accept a lower price for first-period contracts, the fact that $\lambda_i > \lambda_j$ still implies that at any π firms prefer to sign with type i workers at their ask price $r_i^w(\pi)$ rather than with type j at $r_j^w(\pi)$. With this result, it can also be shown that any unraveling of type j workers implies complete unraveling of type i.[12] The order is preserved when buyout and reentry are permitted.

With reentry of firms whose first-period contracts proved unsuccessful, the equilibrium probability that *ex post* productive workers who have not signed first-period contracts receive a payoff 1 in the second-period spot market equals $\pi(0, \ldots, 0)$, just as if no agents had signed first-period contracts. Finding the first-period market equilibrium can be thought of as a pure assignment problem because aggregate uncertainty π is no longer endogenous. Since the ordering property is preserved, the least-promising type of workers who participate in unraveling is type i_* such that the total number of firms is between $\Sigma_{i=1}^{i_*-1} n_i$ and $\Sigma_{i=1}^{i_*} n_i$. There is incomplete unraveling of the cutoff type i_*, and the equilibrium price r_{i_*} of first-period contracts with the cutoff type workers equals their ask price $r_{i_*}^w(\pi(0, \ldots, 0))$. The equilibrium price r_i of contracts with any type i more promising than type i_* is determined by the indifference condition of firms between type i and type i_*:

$$(13)\qquad [1-(1-\lambda_i)\pi(0, \ldots, 0)]v(1-r_i) = [1-(1-\lambda_{i_*})\pi(0, \ldots, 0)] \times v(1-r_{i_*}^w)(\pi(0, \ldots, 0)).$$

Since at any π firms prefer to sign with type i workers at their ask price $r_i^w(\pi)$ rather than with the cutoff type i_* at $r_{i_*}^w(\pi)$, the equilibrium price r_i for type i workers is greater than their ask price $r_i^w(\pi(0, \ldots, 0))$. Moreover, since bid price exceeds ask price for type i_* at

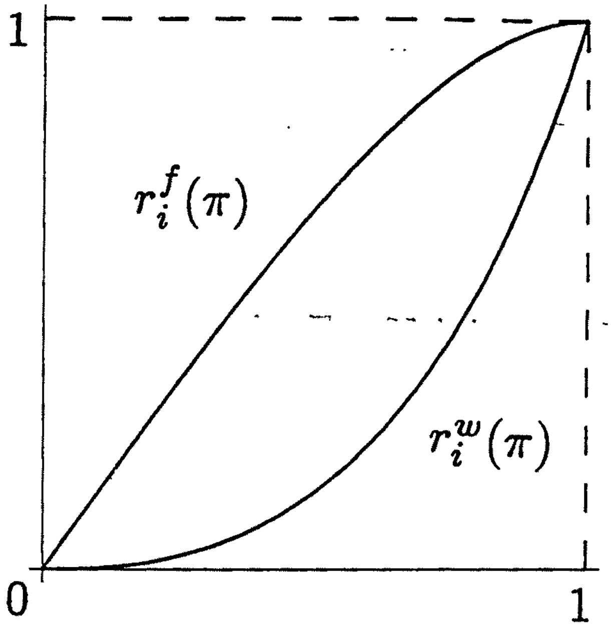


FIGURE 3. ASK AND BID PRICE FUNCTIONS WITH REENTRY

any π, and since firms are indifferent between the cutoff type i_* and any more-promising type i, the equilibrium price r_i of type i workers is below their bid price $r_i^f(\pi(0, \dots, 0))$. Thus, the rent from first-period contracts is shared between workers and firms who sign them.

---

[12] The argument is as follows. Unraveling of type j workers implies that there is a price $r_j \geq r_j^w(\pi)$. Since firms prefer first-period contracts with type i workers at $r_i^w(\pi)$ to contracts with type j at $r_j^w(\pi)$, there exists a price $r_i > r_i^w(\pi)$ such that firms prefer contracting with type i to contracting with type j. Since type i workers also prefer first-period contracts at such r_i to waiting, there is unraveling of type i if there is unraveling for type j. Moreover, unraveling of type i is complete, because incomplete unraveling would imply that $r_i = r_i^w(\pi)$ and firms are indifferent between type i at $r_i^w(\pi)$ and type j at some $r_j \geq r_j^w(\pi)$, which is impossible.

C. *Welfare Comparisons*

Unraveling with recontracting provides some insurance in the first period without sacrificing the *ex post* efficiency of assignments. Therefore, a ban on first-period contracting with reentry hurts market participants as a whole. In particular, firms and workers that are more promising than the cutoff type i_* are strictly worse off under the ban. Welfare of those less promising than the cutoff type is unaffected.

It is also of interest to compare welfare with and without reentry. Reentry shifts bid price functions up because firms are willing to pay more for first-period contracts with any type i of worker at any π. It shifts ask price functions down because type i workers are willing to accept less to sign first-period contracts. More workers and firms sign first-period contracts when reentry is allowed. In general, buyouts must improve welfare, but there is a question of how the gains are distributed among various types of agents. It is possible that some agents become worse off with reentry.

Reentry unambiguously improves welfare of those workers who choose to wait in the equilibrium without reentry. With reentry, the more promising of these workers sign with firms in the first period, and they prefer signing to waiting at $\pi = \pi(0, \dots, 0)$. The rest still wait for the second-period spot market. Both these two groups are better off, because with reentry, $\pi(0, \dots, 0)$ is greater than the equilibrium π without reentry, and so waiting yields a greater expected utility when reentry is allowed. Welfare comparisons for firms and the types of workers who unravel in the equilibrium without reentry are more complicated. Whether allowing reentry makes firms better off depends on the rent they receive from first-period contracts with reentry. Without reentry, firms are indifferent between signing and waiting at an equilibrium π that is smaller than $\pi(0, \dots, 0)$. With reentry, firms prefer signing to waiting at $\pi(0, \dots, 0)$. Allowing reentry makes firms worse off if the equilibrium π without reentry is sufficiently below $\pi(0, \dots, 0)$ that firms are able to sign first-period contracts at very low prices. This occurs when workers are so risk averse for low values of π that they will accept unfavorable terms in first-period contracts without reentry. Similarly, the welfare effects of allowing reentry on the types of workers who participate in unraveling without reentry depend on how reentry changes the rent they receive from first-period contracts.

IV. Concluding Remarks

In labor markets for young professionals and other situations where information is imperfect but matching people to positions is important, uncertainty produces anxiety over how participants will make out and to whom they will be matched. Young people in the marriage market have worried about whether there exists a mate out there meant just for them. Applicants for jobs and for admission to schools are concerned that their preferred firm or school might not want them, and that there may be keen competition for positions. They fear being passed over and not "getting in."

The problem is not confined to applicants. Firms worry about filling available positions. Colleges that fail to attract their targeted freshmen class size leave serious money on the table: the marginal cost of students below capacity is much smaller than average cost per student. Hospitals that fail to find qualified interns and residents, and law firms unsuccessful in recruiting promising associates, face higher costs because other people's time must be diverted to these tasks. Firms behave as if they were risk averse under such circumstances.

The model developed here is the first to analyze the competitive market and pricing aspects of these kinds of problems. Initial uncertainty about each candidate's future productivity is resolved only with time. Everyone, both workers and firms alike, shares common knowledge of these prospects along the way. The tendency toward early contracting—unraveling—noted by Roth and others is a manifestation of attempts to eliminate these kinds of anxieties when, for a variety of reasons, insurance is incomplete. Early contracts provide limited guarantees of positions to parties who accept them. They partially insure workers against being shut out of the market should they prove productive but too many qualified applicants become available later. They partially insure firms against the risks of being shut out of the *ex post* spot market should qualified persons be in short supply. But there are costs of early contracts. They are made with imperfect information. Some of them are bound to be inefficient *ex post*. Nonetheless, the *ex ante* benefits from insurance to participants exceed the *ex post* inefficient matching costs when unraveling occurs. Unraveling is Pareto optimal for those who participate when marketwide insurance is impossible to arrange. In this sense, like all externalities, it is a manifestation or a "missing" market.

A main result of our analysis is that incentives for early contracts are ordered on the productivity potential of applicants. If observed at all, unraveling occurs among candidates who appear to be most promising a priori, before full information is revealed. This result contrasts with models based on private information, where the ordering goes in the opposite direction. For instance, Theodore Bergstrom and Mark Bagnoli (1993) show that in marriage markets where participants possess private information about their prospects, more-promising candidates marry later. Time signals their private information to others. Less-desirable persons do not benefit from waiting and contract early.

There are virtues in both formulations, depending on the application. In real marriage markets, waiting and showing proof of one's superior qualities tends to attract superior mates. Is this not the traditional economic reason why husbands tend to be older than wives in most societies? But in those labor and other markets where unraveling is observed, superior prospects typically are observed to sign early. This is implied by our model. Less-attractive candidates are the ones who wait. In these situations, credentials, test scores, grades in school, and observations of teachers and more-experienced hands are common knowledge. The data are rather more impersonal and more easily communicated directly to outsiders than are traits relevant to familial marriage. They are not the exclusive personal property of the candidates themselves. In fact there are serious questions of whether candidates can even form accurate and unbiased estimates of themselves in these situations (e.g., see Rosen, 1986).

Unraveling is not a common feature of entry-level labor markets. The imperfect information aspects of matching that provoke it are not nearly so important as other features of employment in many labor-market settings. Yet unraveling seems more prevalent today than in an earlier era, when information, standardized tests, and other credentials were far-less routinized. Limitations on the technology of assessing prospects' potential productivity based on common criteria make it more costly to order prospects among participants. When there are more differences of opinion, fears of being shut out *ex post* are of less concern. Standardized tests and other improvements in the market's ability to classify applicants' future potential have encouraged the rush toward early contracts in many situations and likely will continue to do so.

Another important implication of our model is that unraveling does not always occur in a competitive matching market. For unraveling to arise, participants must anticipate that the most-productive workers are likely to be on the short side of the spot market should everyone else choose to wait. This feature also accords with observation.

Unraveling in the medical interns and residents market first occurred after World War II, when there was an extreme shortage of qualified doctors and medical schools, and hospitals were making great efforts to improve their quality. Early competition for clerks among American judges is a very recent phenomena unquestionably associated with the huge increase in demand for judicial litigation and increased court case loads. These demands impose much greater burdens on judges today. Judges caught short or without capable clerks have much more work to do, without a compensating increase in pay, than those who manage to attract talented aides. Similarly, early contracts via summer-intern network connections for young law associates are also a recent phenomenon. It has been caused by the rapid growth of large, general-purpose law firms, in conjunction with the tendency for the size and importance of legal claims to be concentrated on the better firms (Steven Spurr, 1987), and the high concentration of student credentials among a relatively small number of top law schools. Finally, the tendency for young basketball players to skip college or leave early for professional basketball has been associated with a general increase in demand for the game that increased the premium for outstanding talents and the inability or unwillingness of "amateur" teams to compete in the marketplace. Each of these examples lends credence and broad empirical support to our way of looking at the problem.

Given that unraveling is caused by a kind of market failure, it is easy for us as economists to think of improvements. One obvious alternative is for firms to buy out of early contracts that prove unsuccessful *ex post.* Provisions of this kind, however, are seldom observed. Perhaps additional moral hazard considerations are involved. It may be difficult for parties to agree on enforceable tests of what constitutes success. The abolition of mandatory retirement in the United States has made such problems familiar to academics. Or it may be largely distributional elements that suppress it, similar to the group interests that support many inefficient economic policies. The distributional consequences of shutting down the recontract market in our model are less than compelling on that score.

In fact, the phenomena of early contracting and unraveling often provoke efforts to prohibit it, not unlike disarmament treaties. The medical interns and residents matching program that first motivated economic thinking on these matters is such a case. It arose out of the concerted attempts by hospitals to eliminate early contracting and the mismatching that resulted from it. In our model, unraveling, whenever it occurs, unambiguously benefits firms *ex ante* and harms less-promising candidates. Moreover, no compelling case can be made that it harms the better applicants—it is likely to help many of them—so this kind of "collusive" thinking is hard to apply in this context.

Before conceding much on this point, hasten to point that our analysis is limited price mechanisms. An important feature of interns' and residents' matching program that it greatly limits the role of prices. While wages and working conditions vary substantially from hospital to hospital, wages and other terms are posted in advance of the match and applied uniformly to all who are accepted. An analogy would be a situation where each graduate department posted identical fellowship terms that applied to all applicants who are admitted and accepted the offer. The point is that prices are not allowed to clear the assignment market at the actual contract point. Similar restrictions have been observed in allegedly collusive agreements among colleges to limit price cutting in the form of scholarships for freshmen enrollments. While they are outside the scope of our market-clearing model, it certainly seems possible that restrictions on price competition would be preferred by firms.

Finally, aggregate uncertainty about the value of productive characteristics plays a crucial role in our model. There is an important sense in which it causes uncertainty about the availability of positions that motivates the rush to sign early. We believe that these considerations are important for understanding assignment markets for discontinuous types of agents. However, it can be shown (Li and Suen, 1997) that aggregate uncertainty is not necessary for unraveling to occur. Individual uncertainty alone can produce situations where the gain from insurance in early contracts outweighs the loss of *ex post* efficiency from mismatches.

REFERENCES

Bergstrom, Theodore and Bagnoli, Mark. "Courtship as a Waiting Game." *Journal of Political Economy*, February 1993, *101*(1), pp. 185–202.

Crawford, Vincent. "Comparative Statics in Matching Markets." *Journal of Economic Theory*, August 1991, *54*(2), pp. 389–400.

Gale, David and Shapley, Lloyd. "College Admissions and the Stability of Marriage." *American Mathematical Monthly*, January 1962, *69*(1), pp. 9–15.

nt, we
iited to
of the
ram is
While
ubstan-
es and
match
cepted.
re each
fellow-
nts who
he point
the as-
ct point.
ed in al-
colleges
scholar-
they are
clearing
it restric-
preferred

bout the
ays a cru-
important
about the
s the rush
consider-
ng assign-
types of
(Li and
inty is not
Individual
situations
early con-
efficiency

oli, Mark.
." Journal
uary 1993,

e Statics in
f Economic
p. 389–400.
College Ad-
Marriage."
hly, January

Koopmans, Tjalling and Beckmann, Martin. "Assignment Problems and the Location of Economic Activities." *Econometrica*, January 1953, *25*(1), pp. 53–76.

Li, Hao and Suen, Wing. "Risk-sharing, Sorting, and Early Contracting." Working paper, University of Hong Kong, 1997.

Mongell, Susan and Roth, Alvin. "Sorority Rush as a Two-Sided Matching Mechanism." *American Economic Review*, June 1991, *81*(3), pp. 441–64.

Rosen, Sherwin. "Prizes and Incentives in Elimination Tournaments." *American Economic Review*, September 1986, *76*(4), pp. 701–15.

Roth, Alvin. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory." *Journal of Political Economy*, December 1984, *92*(6), pp. 991–1016.

______. "A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom." *American Economic Review*, June 1991, *81*(3), pp. 415–40.

Roth, Alvin and Xing, Xiaolin. "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions." *American Economic Review*, September 1994, *84*(4), pp. 992–1044.

Shapley, Lloyd and Shubik, Martin. "The Assignment Game I: The Core." *International Journal of Game Theory*, 1972, *1*(2), pp. 111–30.

Spurr, Stephen. "How the Market Solves an Assignment Problem: The Matching of Lawyers with Legal Claims." *Journal of Labor Economics*, October 1987, *5*(4), pp. 502–32.

Motivation and Markets

By W. BENTLEY MACLEOD AND JAMES M. MALCOMSON*

Many workers receive pay based on subjectively assessed performance, yet the shirking model of efficiency wages excludes it. This paper incorporates such pay, with the following results. Performance pay is more efficient than efficiency wages when the costs of having a job vacant are low and qualified workers in short supply. More capital-intensive industries pay more than less capital-intensive industries, as observed in studies of interindustry wages differentials. Sustaining an efficient outcome requires a social convention similar to the notion of a fair wage. The model also makes predictions about the relationship between turnover, wages, growth, and unemployment. (*JEL* D82, J41, O15)

The difficulty of measuring individual performance in an objective way limits the use of legally enforceable performance pay to motivate employees. This paper shows that the use of self-enforcing agreements as an alternative can help us understand a number of apparently unrelated phenomena in labor markets. The model developed here has implications that go beyond those of the shirking model of efficiency wages of Carl Shapiro and Joseph E. Stiglitz (1984) commonly applied to such situations. It makes predictions about the form of efficient incentive agreements. It provides a reason for more capital-intensive jobs to offer higher pay. It also provides a reason why notions of fairness may play an important role in the minds of labor-market participants without appearing in econometric studies as an important exogenous determinant of labor-market outcomes.

Efficiency wage models rely on high wages to provide incentives, yet this is not the only form of compensation we observe in practice. Firms also motivate workers with promises of bonuses and promotions. In many cases these promises are only informal or are based on subjective assessments of performance that make them unenforceable at law. For example, in the 1990 Workplace Industrial Relations Survey for the United Kingdom, 34 percent of employees were reported as receiving some form of *merit pay*, "which depended on a subjective judgement by a supervisor or manager of the individual's performance" (Neil Millward et al., 1992). In the 1988–1990 National Longitudinal Survey of Youth (NLSY) for the United States, 14.1 percent of the individuals reported receiving bonuses and 24 percent reported some form of performance pay (a bonus, commission, or piece rate), though we do not know the extent to which this pay is based on subjective assessments of performance. Moreover, employees who are paid fixed wages or salaries may get promoted for subjectively assessed good performance. Hence asking when firms are likely to use subjective rewards is an empirically relevant question.

Clive Bull (1987) derives incentive constraints that ensure firms pay promised bonuses that are unenforcable at law. MacLeod and Malcomson (1989, 1993b) show that one can pool the incentive constraints for the firm and worker and prove that, provided there is sufficient rent from employment, firms may use either efficiency wages (high wages combined with the threat of dismissal) or performance pay (workers paid an end-of-period bonus if their subjectively assessed performance is satisfactory) to motivate employees. This rent is the difference between the returns to the current relationship and those in the spot market for labor.[1] Should either the worker shirk or the firm not pay the bonus, a separation occurs that harms the offending party through the loss of the future rent from the relationship. A practical example of this is the case of the highly paid traders discussed in James B. Stewart (1993) who quit First Boston Bank because their bonuses were smaller than they felt they had been promised, despite receiving total compensation of over one-half million dollars. When there are sufficient rents to the relationship to enforce an agreement, the agreement is called *self-enforcing.*

---

* MacLeod: Department of Economics/The Law School, University of Southern California, Los Angeles, CA 90089; Malcomson: Department of Economics, University of Southampton, Southampton SO17 1BJ, U.K. The authors would like to thank for helpful comments Paul Beaudry, Lorne Carmichael, Janet Currie, Greg Dow, Avner Greif, Bengt Holmström, Stephen Jones, Ashok Kotwal, Thomas Lemieux, and participants in the European Summer Symposium on Economic Theory 1993 (especially David Canning, Robert Gibbons, Dilip Mookherjee, and Michael Riordan) held at the Studienzentrum Gerzensee of the Swiss National Bank. MacLeod is grateful for the financial support of the Social Science and Humanities Research Council of Canada and the National Science Foundation Grant SBR-9709333; Malcomson gratefully acknowledges financial support from the U.K. Economic and Social Research Council (ESRC), award number R000231484.

This paper models the choice between performance pay and efficiency wages to motivate employees as a function of market conditions. In the spirit of F. A. Hayek (1982), we examine the form of compensation that an efficient market would select.[2] With some jobs, the cost of having the position vacant is high relative to the value of a worker's time. Examples are jobs in capital-intensive industries where workers are either on an assembly line or involved in surveying the operation of expensive equipment. If such a position is not filled, the equipment may have to be closed down or else used only inefficiently. At an efficient equilibrium, firms must be able to replace such workers quickly. But if a firm can replace a worker immediately with a close substitute on similar terms, it cannot be earning a rent from employing that worker. Thus the rent required to make the compensation agreement self-enforcing must go to the worker in the form of a high wage, with the threat of job loss to provide motivation. A rent must also go to new employees. If it does not, the firm could increase its current profits by replacing a current employee with a new one at a lower wage. Anticipating this, the current employee would expect no rent from future employment and so shirk. In an efficient equilibrium, therefore, firms pay efficiency wages.

Conversely, consider jobs for which workers are in short supply—for example, sports stars or traders in financial markets. Even though these workers may earn high wages, they could earn high wages at other jobs and hence the threat of firing is not a motivator. In this case the efficient market equilibrium has performance pay. Where performance is subjective and performance pay thus not enforceable in court, firms have an incentive to renege on a bonus at the end of the period. Performance pay can nevertheless be enforced by the threat of workers either quitting or performing poorly in the future, but only if the firm receives a future rent from continued employment. The actions of the disgruntled traders at First Boston described above show that such threats are a practical possibility. In this case, the rent from hiring new employees must also go to firms, as otherwise employees would quit for higher pay even if paid the bonus.

We also consider a simple two-sector economy in which both efficiency wages and performance pay occur in equilibrium. The model predicts that workers in capital-intensive jobs receive higher pay, consistent with the evidence found in William T. Dickens and Lawrence F. Katz (1987) and Kenneth R. Troske (1994). There is then an ambiguous relationship between turnover and pay. In industries that use high wages to motivate workers, relative wages increase with the rate of turnover. Conversely, when workers are in short supply and incentives are provided with bonus payments, total compensation is negatively related to turnover. In the absence of a motivation problem, turnover would not effect equilibrium wages in the model.

Equilibrium in this two-sector economy exhibits some interesting dynamics. We consider the characteristics of the equilibrium when there is a spurt of growth in the capital-intensive sector, resulting in labor moving into that sector and thus increasing turnover in the less capital-intensive sector. This temporary increase in turnover raises the rent required for an agreement to be self-enforcing and results in the short run in a decrease in wages, an increase in unemployment and inequality, and an initial slowdown in total output. All of these effects are temporary, but they signal that the benefits of economic growth to workers may be only long run, not immediate.

---

[1] In the model, firms and workers are risk neutral and hence the total return to a relationship is the sum of profit plus utility in money terms. The *rent* to the relationship is the difference between this sum and the sum of profit plus utility if firm and worker were to take their next-best market alternatives.

[2] It is the role of the market that distinguishes the present paper from MacLeod and Malcomson (1989, 1993b).

The model also provides a way to think about the role of fairness in wage determination. Survey evidence in Roger T. Kaufman (1984), Daniel Kahneman et al. (1986), Alan S. Blinder and Don H. Choi (1990), Truman F. Bewley (1997), and Carl M. Campbell III and Kunal S. Kamlani (1997) indicates that firms believe that treating workers fairly is important. Of the managers interviewed by Blinder and Choi (1990), 79 percent perceived "a wage reduction to take advantage of labor market slack" as unfair, and 95 percent thought their workers would perceive it as unfair. Moreover, 95 percent also thought that a reputation for an unfair wage policy would result in reduced work effort. This illustrates how deeply the convention is ingrained on *both* sides of the labor market. Kaufman (1984) comes to a similar conclusion for Britain, even for nonunionized firms.

In the current model, the level of wages and form of compensation in an efficient equilibrium depend on market conditions, turnover rates, and vacancy costs, and not on any exogenously imposed notions of fairness. However, the existence of a self-enforcing agreement that sustains equilibrium depends on a convention such as fairness. When employees are paid efficiency wages (and thus receive the rent from employment), firms have an incentive to try to lower wages for the current period because it is high future wages, not high current wages, that provide the incentive to perform. If employees anticipate that firms will succeed, the incentive to perform in the previous period is removed. Sustaining an efficient equilibrium thus requires a convention such as fairness that prevents wages being bid down. In the model this convention is an equilibrium because firms believe wage cuts today will result in workers shirking in the future, in exactly the manner described in the survey evidence. Conversely, when bonus payments are used, firms are induced to pay the bonus only because of a rent from continuing the employment in the future. Once a bonus has been paid, workers have an incentive to try to raise wages for the current period because it is only the future rent that provides the incentive for firms to pay. But if firms anticipate that workers will succeed, the incentive to pay the bonus in the previous period is destroyed. Thus sustaining an efficient equilibrium in this case requires a convention that prevents wages being bid up.

The structure of the paper is as follows. Section I reviews the theory of self-enforcing agreements and discusses how the existence of such agreements is related to the concept of a fair wage. Section II examines the nature of a market equilibrium in self-enforcing agreements and shows how the form of agreement is related to the cost of having a job vacant. Section III considers the two-sector version of the model. Section III, subsection A, examines the form of interindustry wage variation as a function of capital intensity in the two sectors. Section III, subsection B, presents simulation results of the effect of growth in the capital-intensive sector.

I. Self-Enforcing Agreements

We model employment as a long-term relationship governed not by a formal contract but, following the insight of Ian R. MacNeil (1974), by an informal agreement about how the parties are to behave. At the beginning of each period, worker and firm agree on a compensation package that includes a fixed wage component and possibly an end-of-period bonus. The worker then decides the level of effort to supply, followed by the firm deciding the level of bonus pay. At the end of the period, either party may decide to leave the current relationship. In the spirit of the incomplete contracts literature stemming from Sanford J. Grossman and Oliver D. Hart (1986), we suppose that, though the performance of the worker is observed by the firm, important aspects of performance are not measurable in a way that can be verified in court.[3] Hence the bonus cannot be made a legally binding contract tying pay to performance. An agreement can be made self-enforcing by the worker quitting or shirking in the future, and by the firm firing the worker or cutting the bonus, if the agreement is broken. This section provides a complete characterization of the set of self-enforcing agreements and illustrates the role of a fair wage convention in ensuring the existence of an equilibrium.

---

[3] See MacLeod (1997) for a formal model of Oliver E. Williamson's (1975) insight that bounded rationality implies contract incompleteness.

Consider a market for homogeneous labor with discrete periods that correspond to the minimum time a worker is employed. There are L identical workers and, in contrast to MacLeod and Malcomson (1989, 1993b), free entry of firms with identical jobs. If there are J_t jobs at time t, the sunk cost of creating an additional job vacancy at t is continuous and nondecreasing in J_t. We denote this cost by $c_t(J_t)$. It consists of the sunk cost element of the capital equipment required for the additional employee and any costs of reorganizing the production process, which may be increasing in J_t because of diminishing returns, congestion costs, etc. This cost must be incurred before the matching process starts for period t so, at the time matching takes place, J_t is fixed.

Employed workers get utility $W_t - ve_t$ from being paid W_t and exerting effort e_t in period t. For simplicity, we assume workers either work ($e_t = 1$) or shirk ($e_t = 0$), with $v > 0$ the disutility of work. The model is more general than this unidimensional interpretation suggests. Not shirking can correspond to the production of a vector of quantitative and qualitative dimensions of performance at an acceptable level. The profit from a job employing a worker in period t is $pe_t - W_t$, where p is the revenue product of labor. Each firm individually takes p as given but it may in aggregate be a decreasing function of the total employment of nonshirking workers in all firms because of a downward-sloping demand curve for the product.[4] Workers without jobs receive utility $u_t > 0$ per period. Thus to retain a worker who always shirks requires a positive wage so, as seems plausible enough in practice, firms never do that.

At the beginning of period t, unemployed workers and firms with vacancies meet in a market to negotiate employment. Firms make offers that workers can either accept or reject. If an offer is rejected, worker and firm reenter the market with expected future utility $\bar{U}_t$ and profits $\bar{\Pi}_t$, respectively. The pay for period t consists of a base wage w_t that the firm pays whatever the worker's performance and a bonus b_t that the firm agrees to pay as long as the worker does not shirk. Consistent with employment at will, the firm is committed to the wage for the current period of employment only. However, for modelling purposes we assume that, at the time of hiring, worker and firm reach agreement (possibly implicit) on wages and bonuses for the life of the relationship. Below we discuss sufficient conditions that ensure the firm does not renege on these future payments.

The wage w_t can be paid either at the start of the period or, as long as the courts enforce it, at the end. (In principle, w_t might be negative, in which case it is equivalent to a performance bond that the worker pays the firm in period t, but this is never required for any of the equilibria we discuss below.) The bonus b_t may be simply a periodic bonus or take the form of a piece rate, commission on sales, crop share, promotion to a higher-paying grade, or any other type of performance related pay based on performance that is not verifiable in court. It thus differs from the fixed wage w_t in that it is paid only at the end of the period and so can be made conditional on the performance of the worker in period t, but can never be legally enforced because courts cannot verify performance.

Once employment for the current period has been agreed, the worker chooses whether to work or to shirk. The firm observes the worker's performance and then decides whether to pay the bonus b_t. At the end of the period, the worker and the firm decide whether to stay matched for the next period. There is a positive probability that they will separate at this point because it becomes unprofitable to continue the job for purely exogenous reasons unrelated to pay. The timing is shown in Figure 1.

---

[4] This could also be because of external decreasing returns to scale, in which case the welfare comparisons discussed later need minor amendments.

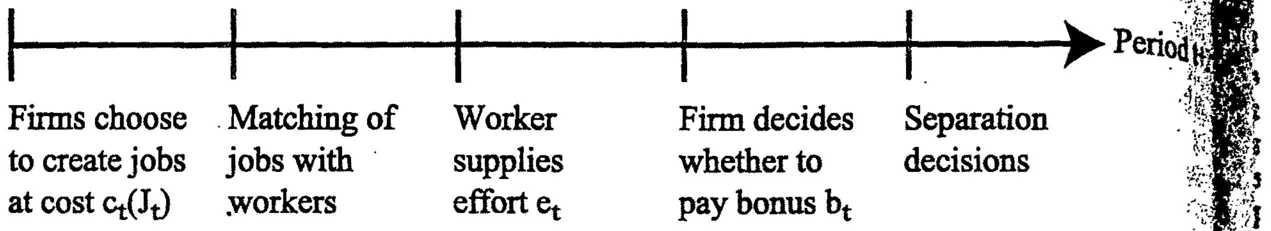

Period
Firms choose to create jobs at cost $c_t(J_t)$
Matching of jobs with workers
Worker supplies effort e_t
Firm decides whether to pay bonus b_t
Separation decisions

FIGURE 1. TIMING OF DECISIONS IN PERIOD t

If neither party reneges, the wage payment in period t is $W_t = w_t + b_t$. Let ρ_t denote the probability that a match at $t - 1$ continues at t if neither party reneges. We assume that ρ_t is independent of how long the match has lasted and, for the moment, that it is less than one only because some jobs become unprofitable for exogenous reasons. For any wage profile, the expected future utility of a worker employed in period t in a match starting at τ if both parties abide by the agreement is

$$(1)\qquad U_t = W_t - v + \delta[\rho_{t+1}U_{t+1} + (1 - \rho_{t+1})\bar{U}_{t+1}], \qquad \text{for all } t \geq \tau,$$

where $\delta \in (0, 1)$ is the discount factor and $\bar{U}_{t+1}$ the expected future utility of an employee whose match terminates at the end of period t.[5] The firm's expected future profit from employing the worker if both abide by the agreement is

$$(2)\qquad \Pi_t = p - W_t + \delta\rho_{t+1}\Pi_{t+1}, \qquad \text{for all } t \geq \tau.$$

(If a job becomes unprofitable for exogenous reasons, the subsequent profit from that job is zero.)

To be self-enforcing, an agreement must be at least as good for both worker and firm as they could get elsewhere in the market. The payoffs must thus satisfy the following individual rationality constraints for the worker (IRW) and for the firm (IRF), with $\bar{\Pi}_t$ the discounted expected future profits from a vacant job:

$$(3)\qquad \text{IRW:}\quad U_t \geq \bar{U}_t, \qquad \text{for all } t \geq \tau,$$

$$(4)\qquad \text{IRF:}\quad \Pi_t \geq \bar{\Pi}_t, \qquad \text{for all } t \geq \tau.$$

In addition it must be in the worker's interest not to shirk and in the firm's to pay any promised bonus. For models of the present type there typically exist many patterns of behavior that are consistent with equilibrium, including ones for which workers are not fired for shirking. It is an insight of Dilip Abreu (1988) that the set of equilibrium payoffs in a repeated game of the present type can be completely characterized in terms of the most severe penalties for cheating that are available. Since our concerns here are primarily with the set of equilibrium payoffs and the form of the wage agreement, we characterize the set of self-enforcing agreements using the most severe punishments. MacLeod and Malcomson (1989) has a more complete discussion.

Given that both worker and firm can terminate the relationship at any time, the most severe penalties are given by the alternatives currently available in the market. If the reasons for a separation were known to the market, it would be possible to establish a reputation with other potential partners for not cheating. We assume here that prospective employers and employees do not know why the previous matches of potential partners came to an end, so establishing an external reputation of this sort is impossible. This is necessarily the case in an anonymous market in which it is hard to keep track of participants, something that may well be true of workers from poor areas large cities in the developing world. But even where it is easier to keep track of participants, the use of reputation depends on there being reliable sources of information about why a separation occurred—the word of mouth of parties to the match is unlikely to be reliable since, if reputation is valuable, neither has an incentive to admit to cheating. Many employers are unwilling to provide information about former employees. In his interviews with (mostly) Connecticut firms, Bewley (1997, Ch. 17 p. 36) found that 69 percent of the 48 employers with whom the issue was discussed reported that it was difficult to get information from former employers. Reasons given included legal problems arising from saying something negative. Only in markets where the reasons for seeking a new match are well known are reputation effects really effective.

---

[5] For notational simplicity, the dependence of utility on the wage profile is not made explicit. Throughout the paper, all payoffs are functions of the agreed wage profile.

Without external reputation effects, $\bar{U}_t$ and $\bar{\Pi}_t$ represent the current spot market alternatives no matter what the reason for separation. Thus to deter shirking by the worker, the expected future utility of keeping to the agreement, U_t, must be at least as great as the utility from shirking (with zero disutility of effort), collecting the wage w_t but not the bonus b_t, being fired, and receiving the discounted future utility $\delta\bar{U}_{t+1}$ from looking for another match at $t + 1$. That is, the wage profile must generate payoffs that satisfy the following incentive compatibility (no-shirking) condition for the worker:

$$(5)\quad U_t \geq w_t + \delta\bar{U}_{t+1}, \qquad \text{for all } t \geq \tau.$$

Substitution from (1) for the left-hand side allows this condition to be rewritten

$$(6)\quad \text{ICW:}\quad E\{\text{future gains to worker}\,|\,t\}$$

$$\equiv \delta\rho_{t+1}(U_{t+1} - \bar{U}_{t+1}) \geq v - b_t,$$

$$\text{for all } t \geq \tau.$$

The intuition for this is as follows. There are two ways to induce the worker not to shirk. One is to ensure that the worker has sufficient gains $\delta\rho_{t+1}(U_{t+1} - \bar{U}_{t+1})$ from continuation of employment in the future that it is worth incurring the disutility of effort in order to avoid being fired even with no bonus. (ρ_{t+1} is the probability that employment continues when the worker does not shirk.) The other is to have a bonus b_t that compensates for the disutility of effort v. If the bonus is sufficiently large then it is not necessary for the worker to earn rents from continuing the relationship. Such payments are explicitly excluded in the shirking model of efficiency wages, which is why it is argued that workers must be offered a wage premium to discourage shirking.[6]

Since the bonus is not legally enforceable, to induce the firm to pay it the expected future profits from paying and having the match continue, Π_t, must be at least as great as from not paying (having already received the product p and being unable to escape paying the wage w_t) and having the worker quit (which results in expected future profit of $\delta\rho_{t+1}\bar{\Pi}_{t+1}$). Firm profits must therefore satisfy the incentive constraint

$$(7)\quad \Pi_t \geq p - w_t + \delta\rho_{t+1}\bar{\Pi}_{t+1},$$

$$\text{for all } t \geq \tau.$$

Substitution from (2) for the left-hand side allows this condition to be written

$$(8)\quad \text{ICF:}\quad E\{\text{future gains to firm}\,|\,t\}$$

$$\equiv \delta\rho_{t+1}(\Pi_{t+1} - \bar{\Pi}_{t+1}) \geq b_t,$$

$$\text{for all } t \geq \tau.$$

The intuition, as in Bull (1987), is simply that the firm will not pay a bonus b_t unless the expected future gains from the employment exceed that bonus.

A worker's threat to quit should the firm not pay a deserved bonus may seem incredible if there is a gain from continued employment. The same applies to a firm's threat to fire a shirking worker. However, MacLeod and Malcomson (1989) show that such threats can form part of a set of self-enforcing social norms. As discussed in more detail below, what is required is for both the worker and the firm to believe that the other would continue to renege if they were to continue the relationship. In more common language, once a worker shirks or the firm refuses a deserved bonus, the relationship sours and both sides believe that it cannot continue, as in the case of the traders at First Boston discussed in the introduction.

---

[6] See Lorne Carmichael (1990) for an extensive discussion of these issues.

The expected future gains to the relationship as a whole are given by the sum of the left-hand sides of ICW and ICF. When summing the right-hand sides, the bonus term cancels to yield the necessary condition for a contract to be self-enforcing that

$$(9)\quad \text{IC:}\quad E\{\text{future gains to relationship}\,|\,t\} \equiv \delta\rho_{t+1}[U_{t+1} + \Pi_{t+1} - (\bar{U}_{t+1} + \bar{\Pi}_{t+1})] \geq v, \quad \text{for all } t \geq \tau.$$

Both worker and firm are risk neutral and share the same discount rate, so the expected future gains to the relationship are independent of the wage profile—the wage terms cancel when (1) and (2) are added.[7] The necessary conditions IRW and IRF depend only on total earnings, not on how these earnings are divided between wages and bonuses. Thus, the individual rationality conditions, IRW and IRF, and the global incentive constraint, IC, are all independent of this division. These three conditions are necessary for the existence of a self-enforcing agreement. MacLeod and Malcomson (1989) show that they are also sufficient. When they are satisfied, it is always possible to find a sequence of wage and bonus payments such that the worker and firm incentive constraints, ICW and ICF, are satisfied.[8]

Since $v > 0$, an implication of the incentive constraint IC is that the existence of a self-enforcing agreement is inconsistent with both worker and firm being indifferent between the current agreement and what they might earn in the spot labor market—IC for period t cannot be satisfied if both $U_t = \bar{U}_t$ and $\Pi_t = \bar{\Pi}_t$. Moreover, this constraint for period t depends only on the *future* returns to the relationship, not on the current wage and bonus payments. It is this fact that necessitates a social norm that may be interpreted as a fair wage. To see this, consider the position in period t of a relationship that started in period $\tau < t$. Suppose, as in the shirking model of efficiency wages, the firm receives no rent, so $\Pi_t = \bar{\Pi}_t$. By IC, the provision of incentives in period $t - 1$ then requires $U_t \geq \bar{U}_t + v/\delta\rho_t$. Thus along any wage profile that provides the incentive for the worker not to shirk, the worker receives utility at the beginning of period t greater than the next-best market alternative $\bar{U}_t$. However, once period $t - 1$ has passed, what is important for incentives at t are any bonus payment for t and the wages and bonus payments from period $t + 1$ onwards, not the wage w_t, so the firm has an incentive to negotiate w_t downwards. It could at the beginning of period t offer a new wage $w_t' < w_t$ that leaves the worker indifferent or slightly better off than at the market alternative *without affecting the incentive to perform in period t*. If, however, the worker anticipates at $t - 1$ that the firm will be successful in reducing the wage at t to the market alternative, the incentives for effort at $t - 1$ will be destroyed.

This problem is solved if the wage w_t has the status of a fair wage. Should the firm attempt to offer at time t a wage w_t' less than w_t, the worker responds by shirking in period t

---

[7] We can see this by defining the *rent* $R_t = U_t + \Pi_t - (\bar{U}_t + \bar{\Pi}_t)$ and substituting from (1) and (2) to give

$$R_t = p - v + \delta(\rho_{t+1}R_{t+1} + \bar{U}_{t+1} + \rho_{t+1}\bar{\Pi}_{t+1}) - (\bar{U}_t + \bar{\Pi}_t).$$

Recursive substitution for future values of the rent results in the wage and bonus terms dropping out.

[8] As written, (9) can be satisfied only if $\rho_{t+1} > 0$, so there cannot be a date t known with certainty to be the last (because, for example, the employee is to retire). Empirically, however, it is a robust finding in experiments with games of this type that players behave in the earlier stages as if the game will continue forever even though it is in fact finite [see Alvin E. Roth (1995 Section III.A.1)]. Moreover, Jacques Crémer (1986) and Michihiro Kandori (1992a) show that, even in theory, incentive compatibility conditions of this type can be satisfied with finitely lived agents if successive generations overlap and acquire information about the play of the preceding generation (as when younger employees in a firm observe the performance and reward of those about to retire).

w_t is greater than the one-period return in the spot market, or by quitting otherwise. Note that it is in the worker's interest to respond in this way if she believes that a firm that succeeds in reducing the wage in this period will also do so in future periods. Responses of this type are in line with the view of 95 percent of the managers interviewed by Blinder and Choi (1990) that an unfair wage policy would result in reduced work effort. Provided the agreed wage profile yields the firm profits equal to or larger than the market alternative, the firm is hurt if the worker responds to a wage cut in this way.[9] Observe that the relationship is reciprocal in the sense that the firm offers a high wage in return for more effort. Reciprocity of this type is consistent with the experimental evidence of Ernst Fehr et al. (1997).

Conversely, consider the case in which the worker receives no rent, so $U_t = \bar{U}_t$. By condition IC, the provision of incentives at $t - 1$ then requires that $\Pi_t \geq \bar{\Pi}_t + v/\delta\rho_t$ and the firm receives profits at t greater than the next-best market alternative $\bar{\Pi}_t$. But once the firm has paid any bonus for period $t - 1$, what is important for incentives are the wages and bonus payments for $t + 1$ on, not the wage w_t. Thus the worker has an incentive to renegotiate the wage upwards to capture some of the firm's rent and if the firm anticipates at $t - 1$ that this will be successful, the incentive to pay the bonus at $t - 1$ will be destroyed. What is required in this case is a social norm ensuring that the worker cannot bid up the wage in this way.

These arguments apply as much to the beginning of the relationship as during it. If in period τ workers in the market believe that a fair starting wage is w_τ, even if the resulting utility is greater than the market alternative ($U_\tau > \bar{U}_\tau$), firms will offer w_τ because they believe that a lower offer would result in shirking workers. Both Kaufman (1984 p. 109) and Bewley (1995 p. 252) are explicit that employers resist lowering pay for new recruits because it would result in their becoming resentful or disgruntled, thus affecting their morale. In this way, paying an above market-clearing wage can be seen as creating the expectation that the worker will reciprocate with high effort. Equally, an offer from an unemployed worker to work for less will be regarded with the suspicion that the worker intends to treat the employment as a way to make a short-term gain by shirking. What is regarded as fair is a convention that coordinates behavior because it results in adverse consequences for those seen to break it, in exactly the same way as the beliefs discussed above. In all sorts of contexts, people who behave unconventionally (by, for example, offering "too good a deal") are regarded with the suspicion that "there must be a catch."

These results provide a link between the problem of worker motivation and the concept of fairness that authors such as George A. Akerlof (1982) have argued are important in the operation of labor markets. The theory adds to the Akerlof model by introducing a set of constraints describing the set of feasible contracts. These constraints form the starting point for the discussion of market equilibrium in self-enforcing agreements in the next section.

II. Market Equilibrium

Given the spot market returns, the constraints (3), (4), and (9) completely characterize the set of wage profiles that can be supported by some self-enforcing agreement. In this section we characterize market equilibria when firms compete with each other by offering self-enforcing agreements. Though the notion of a fair wage is not explicitly discussed, the equilibrium agreements that satisfy the incentive constraints are in each case enforced by an underlying set of self-enforcing norms.

Consider first equilibrium in a stationary environment with a fixed number of jobs J each period and $\rho_t = \alpha$, so there is a constant rate $(1 - \rho_t)$ of exogenous separations that occur because it becomes unprofitable for firms to continue some jobs for reasons other than pay. The cost of creating an additional job is a constant c independent of the number of existing jobs and the revenue product of labor $p(E)$ is strictly decreasing in employment E. To retain stationarity, new jobs are created at a rate equal to the rate of exit, $(1-\alpha)$ per period. In the absence of reputation effects from previous matches, in equilibrium unmatched workers and firms reach agreement immediately and thus $E=\min\{L, J\}$.

---

[9] There are subtle issues to ensure that the threats that we describe are in fact credible. These issues are discussed in some detail in MacLeod and Malcomson (1989), where we show that the threats described here are credible in the sense of being part of a subgame-perfect equilibrium in a well-defined employment game. See also the survey by David G. Pearce (1992) for a review of the issues involved in sustaining cooperation.

In a stationary equilibrium with firms offering the same self-enforcing agreement each period, total pay is $W = w + b$. For any W that is self-enforcing given the market alternatives $\bar{U}$ and $\bar{\Pi}$, there is always a division of pay into a base wage and bonus that satisfies the incentive constraints (6) and (8). Hence it is sufficient to describe an equilibrium state of the economy by (W, J), the total pay for each job, and the number of jobs. Let $U(W, J)$ and $\bar{U}(W, J)$ denote the expected future utility in equilibrium of an employed worker and of a worker entering the labor market, respectively, and $\Pi(W, J)$ and $\bar{\Pi}(W, J)$ the equilibrium expected future profits from a filled job and from a vacant job, respectively. For a stationary equilibrium, it follows from (1) and (2) that worker utility and firm profits from employment under the present assumptions are

$$(10)\quad U(W,J) = W - v + \delta[\alpha U(W,J) + (1-\alpha)\bar{U}(W,J)],$$

$$(11)\quad \Pi(W,J) = [p(E) - W]/(1-\delta\alpha).$$

Recall that $\Pi(W, J)$ is the *ex post* profit from a job after the cost of creating it has been incurred. With separations occurring because jobs become unprofitable with probability $1-\alpha$, the effective discount factor for firms is $\delta\alpha$.

If all workers are employed $(E = L)$, a worker whose job comes to an end can find another straightaway, so $\bar{U}(W, J) = U(W, J)$. Similarly, if all jobs are filled $(E = J)$, a vacant job can always be filled straightaway, so $\bar{\Pi}(W, J) = \Pi(W, J)$. It follows directly that there can be no equilibrium at which the number of jobs J is equal to the number of workers L because that would require $\bar{U}(W, J) = U(W, J)$ and $\bar{\Pi}(W, J) = \Pi(W, J)$. The rent from continuing a match would then be so it is not possible to satisfy the incen compatibility constraint (9) and no se enforcing agreement exists. As in Shapiro Stiglitz (1984), the rent needed for a se self-enforcing social norms is generated by imbalance between demand and supply in labor market. However, as MacLeod Malcomson (1989) observe, this rent can generated by unfilled vacancies just as well by an excess supply of workers. We therefo consider the two possibilities $J < L$ and $J >$ L. When $J < L$ we show that an efficient equ librium involves the use of efficiency wage workers are paid an above market-clearin wage and fired if they shirk. Conversely, whe $J > L$, efficient equilibria involve the use o bonus payments to workers who do not shirk

A. *Efficiency Wage Equilibria*

Consider first the possibility of equilibriu with the number of jobs J less than the numbe of workers L. In this case, $E = J$ and vaca jobs can always be filled straightaway, so the have the same expected future profits as fille jobs. Thus

$$(12)\quad \bar{\Pi}(W,J) = \Pi(W,J) = [p(J) - W]/(1-\delta\alpha)$$

where the second equality follows from (11). To replace jobs that become unprofitable $(1-\alpha)J$ vacancies are created each period. Thu the probability of an unemployed worker finding a job in any period is $\pi(J) = (1-\alpha)J/(L-\alpha J)$. For given (W, J), the equilibrium utilities of employed and unemployed worke are then given by the unique solution to (10) and

$$(13)\quad \bar{U}(W,J) = \pi(J)U(W,J) + [1-\pi(J)][u+\delta\bar{U}(W,J)]$$

where u is the utility during one period of un employment. The utility of both employed a unemployed workers is increasing in W a J.

Two conditions must be met for (W, J) to be a stationary equilibrium. First, firms must be prepared to create enough new jobs to replace those that become unprofitable for exogenous reasons but no more, which requires that $\Pi(W, J)$ equals the cost c of creating an additional job. The equilibrium entry condition is thus

$$(14) \quad [p(J) - W]/(1 - \delta\alpha) = c.$$

Second, neither firms nor workers must be able to gain by cheating on their agreement. Since firms can always fill vacancies, they would always cheat on a bonus. [Formally, with $\Pi(W, J) = \bar{\Pi}(W, J)$, (8) requires that $b \leq 0$ but nothing is lost by setting the bonus to zero.] Thus bonus payments cannot be part of an equilibrium and the only pay is the base wage w. With $b = 0$, the worker's incentive compatibility constraint (6) reduces to the constraint

$$(15) \quad U(W, J) - \bar{U}(W, J) \geq v/\delta\alpha.$$

Since $v > 0$, this constraint implies that employed workers have strictly higher expected future utility than unemployed workers, so the base wage w must be above the market-clearing level. That is, it is an efficiency wage. Substitution of the solutions for $U(W, J)$ and $\bar{U}(W, J)$ derived from (10) and (13) and some tedious algebra (details available on request) allows (15) to be rewritten as the *no-shirking constraint* for workers

$$(16) \quad \text{NSC:} \quad W \geq u + \frac{v}{[1 - \pi(J)]\delta\alpha}.$$

A market equilibrium is a pair (W, J) that satisfies the entry and no-shirking constraints (14) and (16). These constraints are illustrated in Figure 2. The entry condition (14) implies that an equilibrium must lie on the line $W = p(J) - (1 - \delta\alpha)c$ labelled *zero profits*, which is downward sloping because $p(J)$ is decreasing in J. The no-shirking constraint (16) implies that an equilibrium must lie above the upward-sloping line labelled NSC that starts from $u + v/\delta\alpha$ for $J = 0$ and is asymptotic to the line $J = L$. [If there are no jobs, the probability $\pi(J)$ of a worker finding a job is zero and, as the number of jobs approaches the number of workers, $\pi(J)$ approaches 1. This line is also everywhere upward sloping and strictly convex.] Thus the set of efficiency wage equilibria corresponds to the heavy part of the zero profits line to the left of point A in Figure 2.

Shapiro and Stiglitz (1984) identify the efficiency wage equilibrium corresponding to point A. (With their assumption that $c = 0$, the zero profits line can be identified with aggregate labor demand.) However, sustaining even this efficiency wage equilibrium requires a market convention like that found by Blinder and Choi (1990) that prevents firms cutting wages simply because there is excess supply of labor. When one analyzes such conventions formally in terms of self-enforcing social norms in the way done here, point A is not the only efficiency wage equilibrium. Each point in the set of efficiency wage equilibria corresponds to a different fair wage convention. For example, point E in Figure 2 corresponds to all firms offering the wage w_E. Should a firm attempt to enter with a lower wage, workers would interpret it as behaving unfairly and shirk if employed at that wage.

Equilibria with more jobs have higher employment and so exploit more gains from trade. Following Hayek (1982), suppose that competition, experimentation, and evolution result in the market discovering and adopting a set of conventions yielding an outcome that is efficient in the sense that all potential gains from trade are exploited.[10] At the moment there is no generally accepted approach to modelling such a complex process and we do not attempt it here. Rather we begin with the properties of an efficient equilibrium and ask which set of conventions supports such an equilibrium.

For efficiency wage equilibria, efficiency corresponds to maximizing employment. Thus the efficient equilibrium is at point A in Figure 2, the intersection of the no-shirking constraint

---

[10] Essentially this is a notion of *inclusive fitness:* the institutions we observe survive because they help the group survive against competitors. See Robert Boyd and Peter Richerson (1985) for further discussion and additional references.

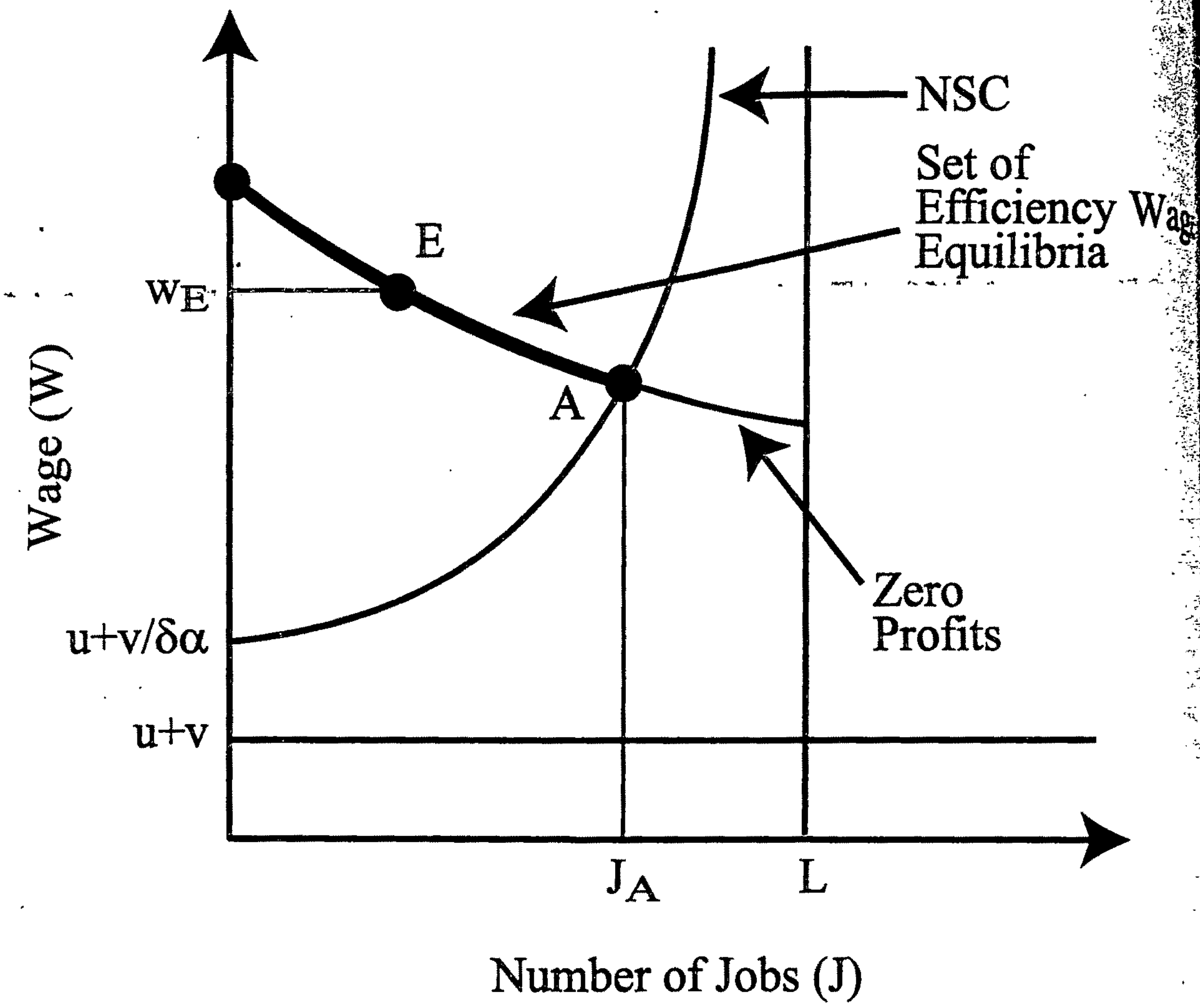


FIGURE 2. MARKET EQUILIBRIA WITH EFFICIENCY WAGES

and the zero profits line, with employment J_A. This is precisely the equilibrium identified by Shapiro and Stiglitz (1984).[11]

How does such a market respond to shocks? Suppose there is an unanticipated but permanent shock that reduces $p(J)$. This shifts the zero profit line downwards. Exactly as i Shapiro and Stiglitz (1984), the no-shirkin constraint is independent of $p(J)$ for given J so the shock shifts point A downwards alon the NSC curve. Thus the new efficient equilibrium has lower wages and employment eve though labor supply is completely inelastic But, if $c > 0$, the move to this equilibrium ca be delayed. Because the cost of creating the J jobs has already been incurred, firms will continue to fill them as long as $p(J_A) \geq W$, thoug they will not create replacement jobs as curren jobs disappear. Indeed, for a shock that reduces $p(J_A)$ by less than $(1 - \delta\alpha)c$, employment and wages need not adjust at all initiall The number of jobs will decline as those th become unprofitable are not replaced until th new efficient equilibrium number is reached

---

[11] Since firms receive zero profits in every equilibrium in Figure 2 and purchasers of the product receive more surplus from higher output, the equilibrium at A will Pareto dominate if workers have higher utility. Workers who would be employed in an equilibrium with lower employment may lose out because the wage is lower, although this is at least partially offset by a higher probability of reemployment when their current job ends. Whether workers on average (or behind the veil of ignorance) gain depends on the cost c and how steeply $p(J)$ declines as J increases.

At that point wages must fall to induce firms to create replacement jobs. How wages evolve in the meantime depends on how market conventions evolve in response to shocks.

B. *Performance Pay Equilibria*

Consider now the possibility of equilibria with more jobs than workers ($J > L$) so that $E = L$. In this case, workers without a job can always get one straightaway and so have the same expected future utility as those with a job. Thus

$$(17) \quad \bar{U}(W, J) = U(W, J)$$
$$= (W - v)/(1 - \delta),$$

the second equality following from use of the first in (10). Employed workers then have the same expected utility as unemployed workers, so efficiency wages are not paid. Exogenous turnover generates $(1 - \alpha)L$ workers seeking new jobs each period. Thus the probability of a vacancy being filled in any period is $\gamma(J) = (1 - \alpha)L/(J - \alpha L)$. For given (W, J), the equilibrium *ex post* profits from filled and from vacant jobs are given by the unique solution to (11) with $E = L$ and

$$(18) \quad \bar{\Pi}(W, J) = \gamma(J)\Pi(W, J)$$
$$+ [1 - \gamma(J)]\delta\alpha\bar{\Pi}(W, J),$$

which gives the expected future profits from a vacancy given the probability $\gamma(J)$ of filling it in each period and the discount factor for firms $\delta\alpha$.

As with efficiency wage equilibria, for (W, J) to be an equilibrium firms must not want to change the number of jobs. Thus the expected post-entry profit per worker must equal the cost of creating a new job:

$$(19) \qquad \bar{\Pi}(W, J) = c.$$

Moreover, neither firms nor workers must gain by cheating on their agreement. Since workers can always get another job, they will shirk unless there is a bonus of amount v to compensate for the disutility of effort, so any equilibrium of this type must have performance pay. [Formally, with $U(W, J) = \bar{U}(W, J)$, (6) requires $b \geq v$ but nothing is lost by setting $b = v$.] Substitution of this into the firm's incentive compatibility condition ICF (8) gives the *no-cheating constraint* for firms

$$(20) \quad \text{NCC:} \quad \Pi(W, J) - \bar{\Pi}(W, J) \geq v/\delta\alpha.$$

Constraints (19) and (20), along with the corresponding efficiency wage constraints for comparison, are illustrated in Figure 3. (Details of derivations are available on request.) The left half of the figure reproduces the curves from Figure 2. For $c > 0$, the zero profit line continues to fall as the number of jobs is increased beyond the number of workers L due to an increase in vacancies and thus a reduction in the probability of finding a worker for the job. (For $c = 0$, it is horizontal to the right of $J = L$). The no-cheating constraint (20) implies that an equilibrium must lie below the upward-sloping line labelled NCC. This line slopes upward because, with more jobs, it takes longer for a firm to fill a vacancy, thus reducing the profits required to prevent reneging on the bonus $b = v$ and increasing what can be paid to workers. It cannot cross the line $J = L$ because with $J = L$ a firm could always replace a worker immediately and would therefore never pay a bonus. The set of performance pay equilibria corresponds to the heavy part of the zero profits line (to the right of point B and with $W \geq u + v$, the lowest wage for which workers will work). In any such equilibrium, firms with jobs that have already been created receive higher profit than they require to continue those jobs. To sustain this requires a convention preventing workers bidding up wages even though there are unfilled vacancies that is exactly symmetric to the convention that prevents firms bidding down wages in efficiency wage equilibria even though there are unemployed workers.

Performance pay equilibria exist if and only if $p(L) - (1 - \delta\alpha)c \geq u + v/\delta\alpha$, efficiency wage equilibria as long as $p(0) - (1 - \delta\alpha)c \geq u + v/\delta\alpha$. Since $p(L) < p(0)$ because of downward-sloping demand for the product, for some parameter values the latter may exist but not the former. The rent per match is decreasing as employment increases, and given that employment under a performance pay

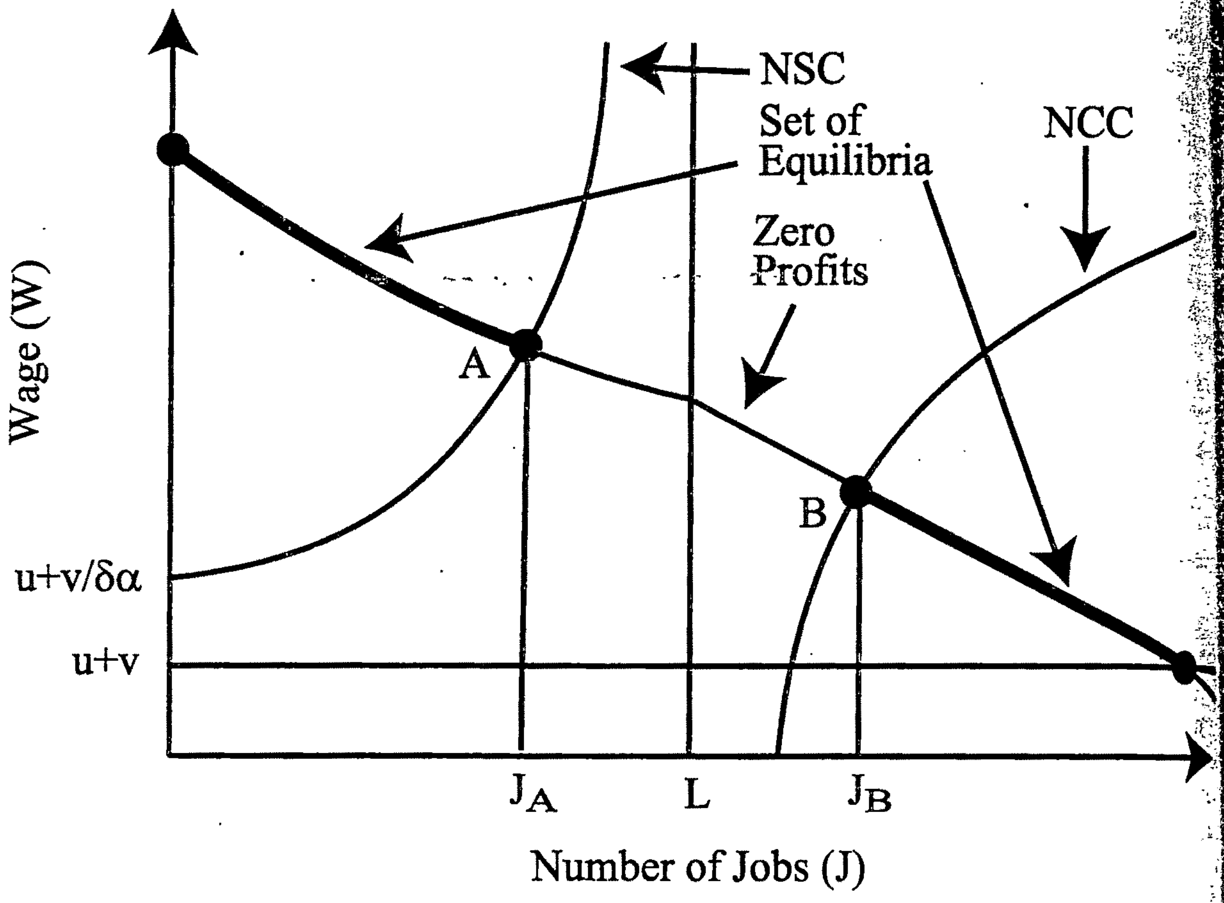


FIGURE 3. MARKET EQUILIBRIA WITH PERFORMANCE PAY AND EFFICIENCY WAGES

equilibrium is higher, it is more difficult to satisfy the incentive constraints at a performance pay equilibrium.

As in the efficiency wage case, individual incentives are not by themselves generally sufficient to determine a unique equilibrium. All performance pay equilibria have the same employment L and the same output. Due to the cost of replacing jobs that become unprofitable for reasons other than pay, an efficient equilibrium minimizes the number of unfilled vacancies. Thus the unique efficient performance pay equilibrium is that denoted by B in Figure 3 with J_B jobs. It Pareto dominates other performance pay equilibria since output and employment are the same, wages are higher, and firms receive zero profits in all such equilibria.

Productivity shocks affect the efficient performance pay equilibrium in a particularly simple way. It can be shown that a reduction in $p(L)$ reduces both the NCC and the zero profits lines by the same amount for given J. In consequence an unanticipated but permanent fall in $p(L)$ leaves the efficient number of jobs unchanged but requires a fall in total pay. Thus pay responds to such a shock but employment and vacancies remain unchanged unless the shock is sufficiently adverse that a performance pay equilibrium no longer exists. If it is sufficiently adverse, there must be a transition to an efficiency wage equilibrium with a consequently large fall in employment, but the dynamics of how that happens depend on how market conventions evolve in response to shocks.

C. *Efficiency Wage versus Performance Pay Equilibria*

How does the performance pay equilibrium with J_B jobs compare with the efficiency wage equilibrium with employment J_A? The cost of moving from the efficiency wage equilibrium with J_A jobs to the performance pay equilib rium with J_B jobs and employment L is the cost of creating the additional jobs which, amortized on a per period basis, is $(1 - \delta\alpha)(J_B - J_A)c$, plus the additional disutility of work $(u + v)(L - J_A)$. The benefit from the move is the benefit from the additional output from increasing employment from J_A to L. Thus there is a net gain from the move if and only if

$$(21) \quad \int_{J_A}^{L} p(J)\, dJ > (1 - \delta\alpha)(J_B - J_A)c + (u + v)(L - J_A).$$

When it is costless to create a job ($c = 0$) as in Shapiro and Stiglitz (1984), the performance pay equilibrium is certainly more efficient whenever the value of employment $p(L) - v$ is greater than the gain from unemployment u, the same condition (marginal product greater than disutility of work) for it to be efficient to employ all workers even if there were no problem of contracting on effort. If either the cost of creating a job is sufficiently high or the marginal product $p(J)$ falls sufficiently over the range J_A to L, the efficiency wage equilibrium is more efficient.

One result from the analysis is that, while efficiency wage equilibria have unemployment, performance pay equilibria do not. This points to a potentially testable prediction. Bonus payments in the model are different from other forms of incentive pay such as piece rates in being in effect discretionary, not contractually fixed. Previous work, such as Charles Brown (1990), has focused on pay systems that link pay formally to some performance measure. Yet in many occupations it is common practice to make performance pay discretionary, as with the traders at First Boston discussed in the introduction. The model predicts that if the unemployment rate for an occupation is low, we should observe the use of such bonus payments, while in a high unemployment environment we should observe efficiency wages. Note that this prediction does not depend on markets locating at one of the efficient equilibria A or B.

Unfortunately standard data sets such as the Panel Study of Income Dynamics (PSID) and the National Longitudinal Survey of Youth (NLSY) do not distinguish between discretionary payments and contractually set incentive pay.[12] One data set that does is the 1990 British Workplace Industrial Relations Survey. Though it does not report the size of discretionary merit pay, it does explicitly ask if a worker received merit pay, as opposed to performance pay with "a mechanical relationship between the worker's pay and some relatively objective measure of output" [Neil Millward et al. (1992 p. 258)]. The above analysis suggests that, in data grouping workers in different labor markets, the incidence of merit pay should be higher for groups with a lower unemployment rate. In Figure 4 we plot the incidence of merit pay by broad occupation against the occupational unemployment rate taken from the 1990 British Labour Force Survey (with the line representing the OLS fit that has slope −2.81 and t-ratio −15.95).[13] The relationship is negative, as predicted. Note that the incidence of merit pay varied from 10 percent to 35 percent. Of course this is not a test of the theory. That would require better data and a carefully specified econometric model. It does, however, illustrate that merit pay is a phenomena that is empirically relevant, as well as one about which theory can make predictions.

The property that equilibria in the model have all rent going to the long side of the market (to workers if there is unemployment, to firms if there are unfilled vacancies) applies to new matches even if they are not required to offer the same agreements as current matches and also to nonstationary environments. The essential intuition is that a firm or an employee

---

[12] For some years these data sets have questions that distinguish between piece rates, commissions, and bonus pay. MacLeod and Daniel Parent (1997) suppose that the incidence of bonus pay is a noisy measure of the extent of subjective performance pay and also find a negative relationship between unemployment and the incidence of bonus pay.

[13] The broad occupational groups are professional and managerial; clerical, administrative, and secretarial; supervisors; skilled manual; semiskilled manual; and unskilled manual.

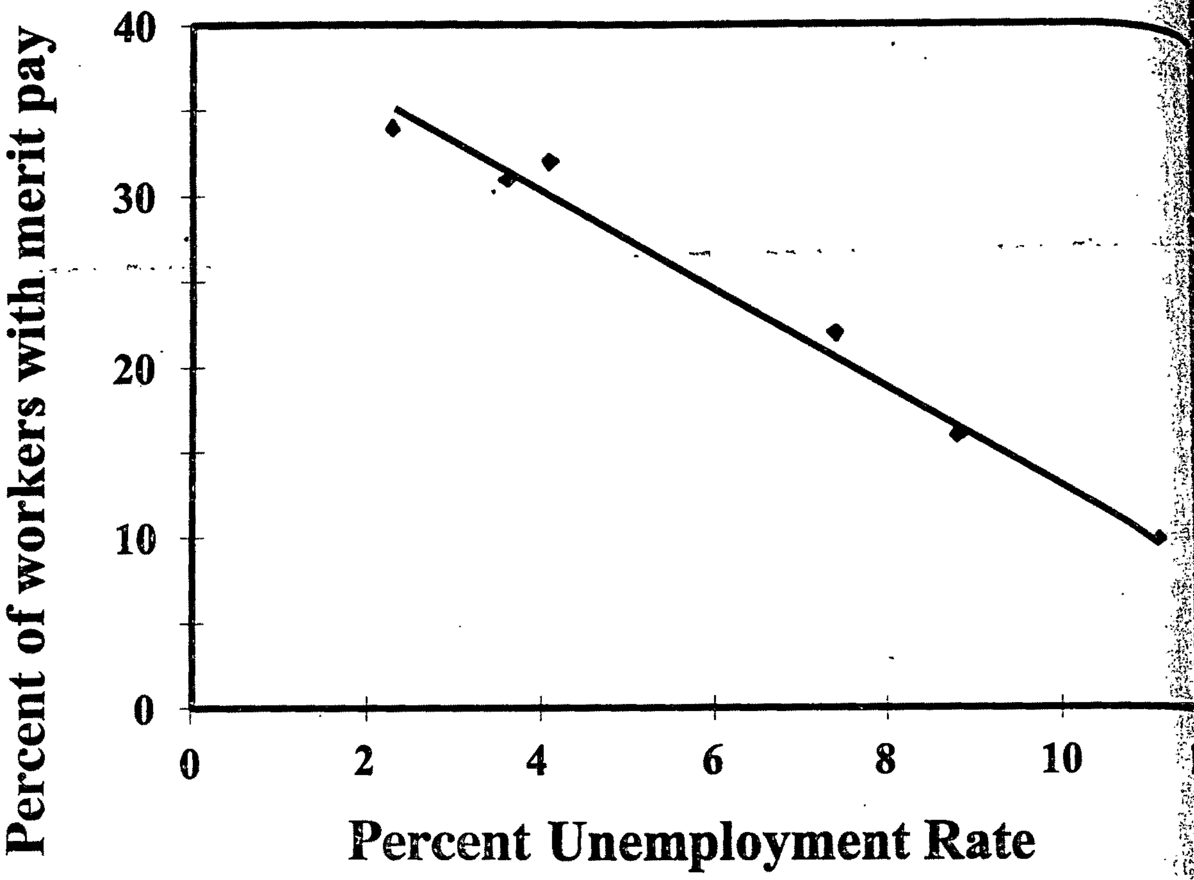


FIGURE 4. PERCENTAGE OF WORKERS RECEIVING MERIT PAY VERSUS UNEMPLOYMENT RATE BY OCCUPATIONAL GROUP

receiving no rent from a current match cannot receive a rent from a new match without making it profitable to renege on its current agreement. The details are in the Appendix. With nonstationary equilibria, however, the same distribution of rent at the start of a match can be achieved with different time paths of payment and thus with different firms having different forms of contract with their workers. Thus different forms may coexist, as found by Brown (1992) for production workers in very narrowly defined industries, though even for narrowly defined industries it is hard to be sure that this is not the result of unobserved heterogeneity among employees. MacLeod and Malcomson (1993a) have further discussion of nonstationary equilibria.

In practice there may, of course, be matching frictions that mean even those on the short side of the market cannot guarantee to find another match straightaway. As long as there is some exogenous turnover, there will then necessarily be both unemployed workers and filled vacancies.

III. Equilibrium in a Two-Sector Econom

A. *Capital Intensity and Wages*

This section illustrates in a simple tw sector economy how the relative capital inte sity of jobs has implications both for the le of wages and for the relationship betw wages and turnover. A number of studies, cluding Alan B. Krueger and Lawrence Summers (1988), John M. Abowd et (1993), and David Card and Krueger (19 suggest that there are variations in wa across markets and industries that cannot explained by standard competitive mode Dickens and Katz (1987) find that high wa are associated with more profitable capi intensive industries. A simple rent-sharing gument would suggest that *all* workers in s firms earn higher wages. However, work by Steve J. Davis and John Haltiwanger (1991) finds that production workers, rather than nonproduction workers, account for much of the interindustry wage variation. Using matched worker-firm data, Troske (1994) finds that much of this interindustry wage variation for production workers can be explained by the capital intensity of the job.[14]

We capture the distinction between relatively capital-intensive and relatively labor-intensive jobs in its starkest form with the following assumptions. Capital-intensive jobs are characterized by a limited number J_t of positions in period t that can be increased only at a sunk cost of $c_t(J_t)$ per job as discussed in the previous section. We assume $c_t(J_t)$ increases sufficiently with J_t as a result of diminishing returns, congestion costs, etc., that the number of production jobs is bounded and less than the number of available workers. Each capital-intensive job has productivity (net of nonlabor costs) of p^k independent of the number of jobs filled. In contrast, there is no sunk cost to creating additional labor-intensive jobs so, even in the short run, there is a potentially unlimited number of such jobs. Assuming an unlimited number simplifies the analysis without affecting its qualitative features—the essential characteristic is that workers are on the short side of the market in this sector. Each labor-intensive job has productivity (net of nonlabor costs) of $p^\ell < p^k$. To highlight the effect of capital intensity, we assume both sectors face exactly the same problem of worker motivation.

Workers are freely mobile between the two types of jobs, so the probability ρ_t^i that a job of type i continues in period t must now reflect not only exogenous separations but also quits to the other type of job. Free mobility implies that workers seeking jobs have the same utility no matter what type of job they seek, so $\bar{U}_t^k = \bar{U}_t^\ell$.

Under these conditions, an efficient equilibrium has all workers in the labor-intensive sector employed in every period. Thus these workers can always find a labor-intensive job, implying $\bar{U}_t^\ell = U_t^\ell$. Together with free mobility, this implies

$$(22) \qquad U_t^\ell = \bar{U}_t^\ell = \bar{U}_t^k.$$

The incentive condition IC (9) then implies that labor-intensive jobs with a worker must yield a rent to firms in period t.[15] With an unlimited number of such jobs, the expected profits from jobs without a worker are zero ($\bar{\Pi}_t^\ell = 0$). Together these imply $\Pi_t^\ell = v/\delta\rho_t^\ell$. If performance was contractible, workers would receive all this rent in pay and hence equilibrium pay would be $W^* = p^\ell$. Use of these in the expression for profits (2) in Section I implies that equilibrium pay in labor-intensive jobs is

$$(23) \qquad W_t^\ell = W^* - v[(1 - \delta\rho_t^\ell)/\delta\rho_t^\ell].$$

Thus, as compared with the situation with contractible performance, workers give up a reward of $v[(1 - \delta\rho_t^\ell)/\delta\rho_t^\ell]$ each period that ensures the firm has an incentive to pay the end-of-period bonus should the worker perform well.

Because capital-intensive jobs have higher productivity than labor-intensive jobs, it is efficient to have all these jobs filled every period. In equilibrium, capital-intensive jobs can thus fill vacancies immediately and so, in contrast to labor-intensive jobs, firms do not face a loss in profits when a worker leaves. Therefore $\Pi_t^k = \bar{\Pi}_t^k$.[16] In this case, to ensure that the incentive constraint IC (9) is satisfied, employed workers must earn a rent $v/\delta\rho_t^k$ from employment and hence

---

[14] See Michael Kremer (1993) and Kremer and Eric Maskin (1996) for an alternative competitive explanation for these observations. See Martin L. Weitzman (1989), Kevin Lang (1991), and James D. Montgomery (1991) for alternative explanations based on recruitment strategies under incomplete information.

[15] When turnover occurs because workers quit, not just because jobs become unprofitable, equations (2) and (7) in Section I need some amendments when $\bar{\Pi}_{t+1} \neq 0$, but these amendments cancel out in deriving (8) and (9).

[16] In practice there are always some costs to filling a vacancy, such as training. In this section we are concerned with search costs arising from the time it takes to find a suitable replacement.

$$(24)\qquad U_t^k = \bar{U}_t^k + v/\delta\rho_t^k,$$

$$= U_t^\ell + v/\delta\rho_t^k.$$

Here the second line follows from (22).

Expressions (23) for pay in labor-intensive jobs and (24) for utilities in the two sectors can be used in the equation for worker utility (1) from Section I, together with (22), to show that pay in capital-intensive jobs is

$$(25)\quad W_t^k = W_t^\ell + v\left(\frac{1-\delta\rho_t^k}{\delta\rho_t^k}\right),$$

$$(26)\qquad = W^* + v\left[\left(\frac{1-\delta\rho_t^k}{\delta\rho_t^k}\right) - \left(\frac{1-\delta\rho_t^\ell}{\delta\rho_t^\ell}\right)\right].$$

Equation (25) relates the size of the wage differential to the rate of turnover in capital-intensive jobs $(1-\rho_t^k)$, the disutility of effort v, and the discount factor δ. As the rate of turnover decreases, so does the wage differential $(\partial(W_t^k - W_t^\ell)/\partial\rho_t^k < 0)$. The wage differential does not depend on the turnover rate for labor-intensive jobs. For labor-intensive jobs, in contrast, it follows from (23) that wages increase as turnover falls $(\partial W_t^\ell/\partial\rho_t^\ell > 0)$. Thus the relationship between turnover and wages is different for different types of job. In sectors of the economy where bonus payments are observed, the theory predicts a negative relationship between turnover and wages and conversely when workers are paid a fixed wage.

The essential characteristics of capital-intensive jobs in the model are that: (i) they have higher productivity, and (ii) there is a sunk cost to creating vacancies that cannot be recouped if the vacancy is not filled. This fits better with the descriptions of Ford's $5-day example in Daniel M. G. Raff and Summers (1987) and Raff (1988) than does the shirking story in Jeremy I. Bulow and Summers (1986), in which efficiency wages are higher where monitoring is more difficult. Raff (1988) comments that the shift to assembly-line work that Ford had adopted must [illegible] made at least some aspects of monitoring easier (for example, speed of work) and that there was actually an increase in the ratio of supervisory to production employees in the aftermath of the introduction of the $5 day. [illegible] Raff (1988 p. 395) notes that "the technical change that made monitoring easier also involved extraordinarily highly dedicated physical capital and a production process that sent unprecedented numbers of pieces past each worker in each unit of time." Thus the cost to Ford of having a workstation not effectively operated would have been high. It is precisely this cost of not having each position filled all the time that makes it efficient in the present model for the market to allocate the rent required for a self-enforcing agreement to employees—for jobs to have a rent requires them to remain unfilled for a time whenever a match ends.

The analysis applies to both the case in which workers can search for higher-paying capital-intensive jobs while employed in labor-intensive jobs and the case in which they must quit their labor-intensive job before they can search for a capital-intensive one. There are two main differences between these cases: (i) the amount of unemployment, and (ii) the labor turnover rate in labor-intensive jobs. When workers in labor-intensive jobs can search while employed, there is no unemployment because, if unsuccessful in getting a capital-intensive job in one period, they can always continue in their labor-intensive job without prejudicing their chance of getting a capital-intensive job in the next. When they cannot search on the job, those who are unsuccessful in getting a capital-intensive job remain unemployed for that period. In this case, the model generates unemployment in the capital-intensive sector, just as in the urban sector of the model of John R. Harris and Michael P. Todaro (1970), despite wages being determined endogenously rather than specified exogenously. Moreover, unlike in the reformulation of that model due to Karl Ove Moene (1988), the wage differential does not depend on differences between the two sectors in the ability to monitor performance, only on differences in capital intensity.

B. Equilibrium and Growth

In this section we simulate some dynamic properties of the model for the Harris-Todaro case in which workers must quit jobs in one sector to look for jobs in the other. In this case labor-intensive jobs may be interpreted as employment in rural areas, in agriculture, or in a small enterprise involving a minimal amount of capital investment. Capital-intensive jobs may be interpreted as the urban manufacturing sector.

Because productivity in capital-intensive jobs is higher, all those jobs are filled in an efficient equilibrium and, as discussed in Section III, subsection A, they have higher pay than labor-intensive jobs. With search on the job not possible, workers enter the capital-intensive sector by moving into the pool of the unemployed. In our simulation we consider the effect of increasing the number of capital-intensive jobs and the impact that the resulting migration from the labor-intensive sector has on output, unemployment, wages, and inequality.[17]

In both sectors some labor turnover occurs because jobs become unprofitable for exogenous reasons. The probability of a match being terminated in each period for such reasons is denoted $(1 - \alpha^i) \in [0, 1)$, for $i \in \{k, \ell\}$. Because the number of capital-intensive jobs is increasing and they have higher pay, there is no reason for endogenous turnover in this sector, so $\rho_t^k = \alpha^k$ for all t. Migration is independent of exogenous separation, so the probability ρ_{t+1}^ℓ that a labor-intensive match that exists in period t continues in period $t + 1$ is

$$(27) \qquad \rho_{t+1}^\ell \equiv \alpha^\ell \left(\frac{L_{t+1}^\ell}{L_t^\ell} \right),$$

where L_t^ℓ is the number of workers in the labor-intensive sector (which is nonincreasing over time due to the growth in the number of capital-intensive jobs).

The time path of workers in the labor-intensive sector, $\hat{L}^\ell \equiv \{L_t^\ell\}_{t=0}^\infty$, determines the rate of turnover from (27) and hence the wage rate each period from expression (23). Because moving to the capital-intensive sector requires joining the pool of unemployed with the same expected utility in equilibrium as staying in the labor-intensive sector, the equilibrium utility in the labor-intensive sector as a function of the time path of workers in that sector is

$$(28) \qquad \bar{U}_t^\ell(\hat{L}^\ell) = U_t^\ell(\hat{L}^\ell) = \sum_{j=t}^{\infty} (W_j^\ell - v)\delta^{j-t};$$

$$(29) \qquad = \sum_{j=t}^{\infty} \left(p^\ell - \frac{vL_{j-1}^\ell}{\delta\alpha^\ell L_j^\ell} \right) \delta^{j-t},$$

the first equality following from (22), the last from (23).

In the capital-intensive sector, there are $L_t^k \equiv L - L_t^\ell$ workers at the beginning of period t and $\alpha^k J_{t-1}$ jobs remaining filled from the previous period. Thus the probability π_t that an unemployed worker gets a job in period t is given by the number of vacancies $(J_t - \alpha^k J_{t-1})$ divided by the number of unemployed job seekers $(L_t^k - \alpha^k J_{t-1})$. Hence,

$$(30) \qquad \pi_t = \frac{J_t - \alpha^k J_{t-1}}{(L - L_t^\ell) - \alpha^k J_{t-1}}.$$

The default utility for workers entering the queue for capital-intensive jobs is then

$$(31) \qquad \bar{U}_t^k(\hat{L}^\ell) = \pi_t U_t^k(\hat{L}^\ell) + (1 - \pi_t) \times [u^k + \delta \bar{U}_{t+1}^k(\hat{L}^\ell)],$$

where u^k is the utility received from one period of unemployment in this sector. Together (24), (29), (30), and (31) determine the expected lifetime utility of employed and unemployed workers as functions of the numbers of workers in the labor-intensive sector, and jobs in the capital-intensive sector, in each period. The time path for workers in the labor-intensive sector as a function of the number of capital-intensive jobs can then be solved using the equilibrium condition that workers migrate until they are indifferent between sectors

---

[17] Formally, because the number of capital-intensive jobs is endogenous to the model, the increase should be seen as the result of a fall in the cost of creating those jobs.

$$\bar{U}_t^k(\hat{L}^\ell) = \bar{U}_t^\ell(\hat{L}^\ell), \qquad \text{for all } t. \tag{32}$$

The effects of increasing the number of capital-intensive jobs are simulated for rapid growth that takes the capital-intensive sector from 20 percent of the workforce to about 80 percent over a period of slightly less than ten years. The productivity of capital-intensive jobs is assumed to be 50 percent higher than that of labor-intensive jobs, so the productive potential of the economy grows about 25 percent in the ten years. The parameter values used in the simulation are: $v = 1$, $\delta = 0.91$, $\alpha^k = \alpha^\ell = 0.9$, $p^k = 4.5$, $p^\ell = 3$, $u^k = 0$.[18] Figure 5 depicts the effect of this growth on unemployment, output (GNP), wages, and inequality of worker incomes as measured by the Gini coefficient. The dashed line represents the number of capital-intensive jobs at each date. In the short run the unemployment rate rises dramatically from about 7 percent to almost 25 percent at its peak, while wages in capital-intensive jobs fall by about 20 percent. After rising slowly, GNP dips slightly and there is a marked increase in inequality. But then GNP starts to rise sharply and, in the long run, wages and the unemployment rate return to their pregrowth levels. The effect of growth is to slightly raise inequality in the long run.

In the simulations, increased turnover in labor-intensive jobs as workers move to the capital-intensive sector lowers the value of a match to firms with labor-intensive jobs. Consequently, pay must fall to ensure sufficient rent to firms in that sector for the wage agreement to be self-enforcing. This reduces utility in labor-intensive jobs. In equilibrium, the expected utility from search in the two markets must be equal, so growth also decreases utility in the capital-intensive sector in the short run. This is achieved by short-term reductions both in pay and in the probability of an unemployed worker getting a capital-intensive job. Note that we do not *assume* workers have to spend some time unemployed after migrating to capital-intensive sector before they can capital-intensive job. Lucky migrants get without being unemployed at all. The creased unemployment is simply what i quired to keep the efficient equilibrium pa conditions satisfied.

Once growth in the number of capi intensive jobs begins to slow, migration f and wages start to increase. Once gro stops, the unemployment rate in the capi intensive sector and real pay in both sec return to their original values. Growth can increase pay in either sector as long as th are workers in labor-intensive jobs to be dr into the capital-intensive sector since labor-intensive sector has constant returns scale. However, those migrants who succ in getting capital-intensive jobs earn hig wages than workers who remain in the lab intensive sector, so average wages incre and there is also an increase in inequality.

If growth in the number of capital-intens jobs continued indefinitely, there wo eventually be no workers employed in labor-intensive sector, that sector would appear, and wages in capital-intensive j would start to rise. Wages could rise befo this if technological progress increased pr ductivity in the labor-intensive sector becau by (23), W_t^ℓ increases by the same amount p^ℓ and that feeds through to wages in capit intensive jobs. This is a source of real wa growth we have ignored in the simulation th could ameliorate, and possibly outweigh, absolute decline in real wages resulting fr increased migration.

As in Harris and Todaro (1970), migrat and urban unemployment are the con quences of growth in the limited number jobs in the high wage manufacturing (capit intensive) sector in urban areas. The import difference is that here we have modelled reasons for the wage differential between ban and rural areas and this allows us to mo the consequences of urban growth for wag Such growth, as we have shown, can result a fall in real wages in both sectors and so, least in its early stages, in a lowering of liv standards for all workers except those who grate and are successful in getting one of increased number of urban jobs. In con

---

[18] The simulation was programmed in Matlab. The code used is available upon request.

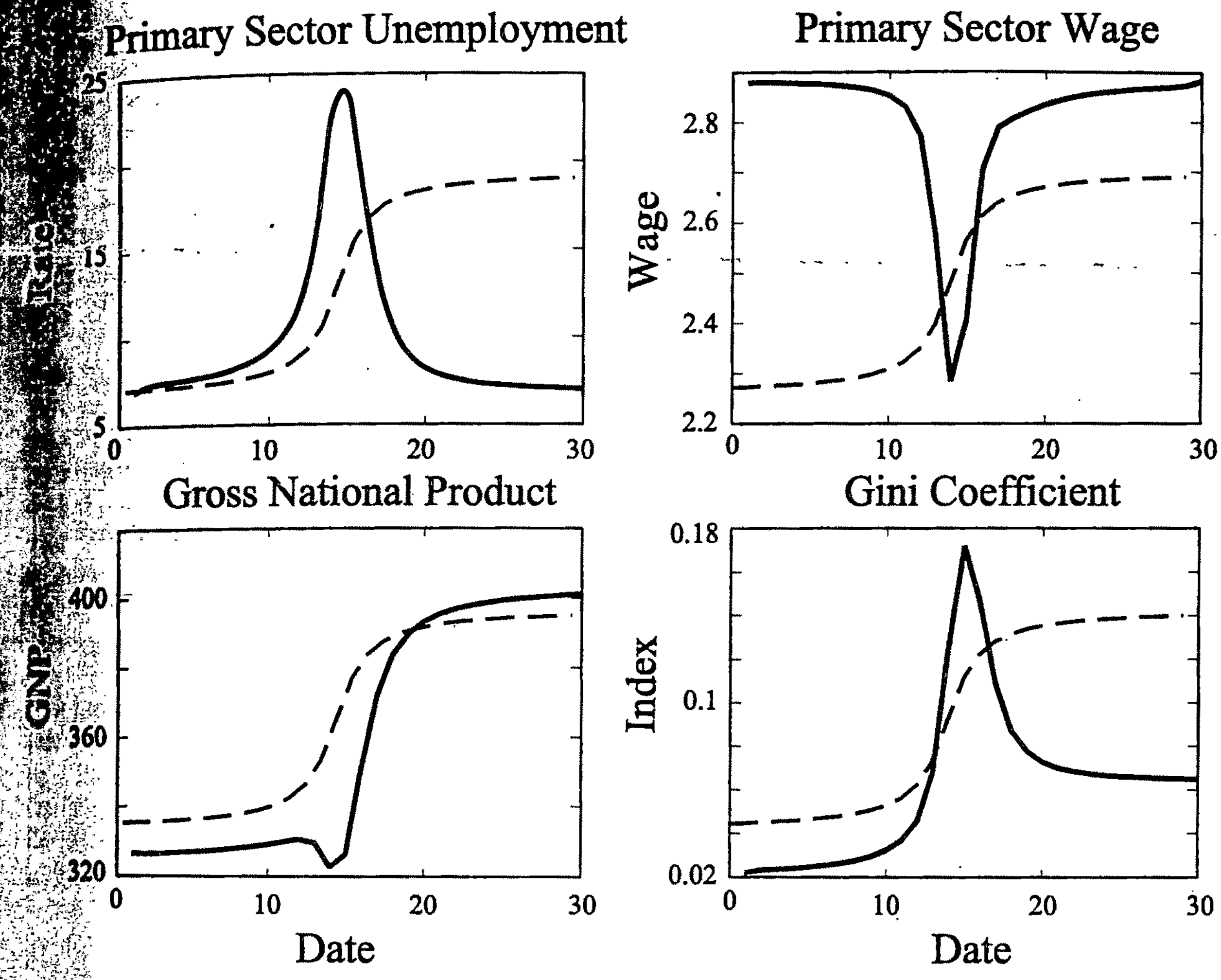


FIGURE 5. ECONOMIC CONSEQUENCES OF GROWTH

quence, the share of profits increases and, if profits go to the already better off, growth generates the Simon Kuznets (1955) effect of an increase in the inequality of the income distribution.

IV. Conclusion

Shirking models of efficiency wages suppose that firms motivate workers by paying above market-clearing wages and threatening to fire workers who shirk. Such models explain neither the widespread use of performance-related pay based on subjective assessments of performance nor the social norms surrounding the concept of a fair wage. We have shown that the theory of self-enforcing agreements can extend the basic efficiency wage model to incorporate both these phenomena.

The model developed here predicts that efficiency wages are likely to be observed in industries where the cost of having a job vacant is high relative to the cost of a worker remaining unemployed (or underemployed). We show that the existence of an efficiency wage equilibrium depends on a set of self-enforcing social norms that ensure firms do not cut wages purely to take advantage of an excess supply of workers, though wages still fall in response to an *increase* in excess supply. Where there is excess demand for workers, the model predicts that firms use some form of subjective performance pay to motivate employees. In that case, the threat of a worker quitting ensures firms have an incentive to pay the promised bonuses. See MacLeod and Parent (1997) for some empirical evidence consistent with these predictions.

Applied to a two-sector version of the model, these results imply that intersectoral wage differentials can arise from differences in the capital intensity of jobs. Moreover, a period of rapid growth in capital-intensive jobs may generate an initial fall in output, reduced wages, higher unemployment, and increased inequality—results that do not follow from standard textbook models.

The model presented here is highly stylized. It does, however, illustrate that viewing the problem of worker motivation from the more general perspective of self-enforcing agreements in a competitive market generates implications that are different from other incentive models. There are many ways in which this approach can be developed. Gary Ramey and Joel Watson (1997), for example, have explored the consequences of incorporating specific investments with sectoral shocks into a model with self-enforcing contracts.[19] Further extensions would include incorporating variations in worker ability and search costs that should result in new and interesting testable hypotheses.

APPENDIX: NONSTATIONARY EQUILIBRIA

This Appendix shows that the property that equilibria in the model have all the rent going to the long side of the market (to workers if there is unemployment, to firms if there are unfilled vacancies) applies to all equilibria, not just the stationary equilibria discussed in Section II. The absence of reputation effects from previous jobs ensures that all agents on the short side negotiating new matches in period t receive the same payoff—because they can always find another match, they will refuse any offer that gives a payoff lower than the best currently available. Thus, if there are more workers than jobs, firms can always fill vacant jobs, so their future profits if their match ends are just the future profits from forming a new one. Formally, if $L > J_t$, then $\bar{\Pi}_t = \Pi_t(\hat{W}_t)$, where $\hat{W}_t$ is the pay profile negotiated matches started at t. Conversely, if there more jobs than workers, workers can ways find another job, so their future util if their match ends is just the future util they would get from forming a new o That is, if $L < J_t$, then $\bar{U}_t = U_t(\hat{W}_t)$. Th properties follow simply from the absen of external reputation effects and all sho side agents being matched.

To satisfy the incentive compatibility (I condition (9) requires a strictly positive re of at least $v/\delta\rho_t$ for matches formed *before* (At $t = 1$, the first period the market operat there are no existing matches for which ince tive compatibility must be satisfied.) That, i turn, implies the same rent for new match started at $t > 1$. This can be seen as follo In adding (1) and (2) for any τ, the pay te cancel, which implies $U_t(\hat{W}_\tau) + \Pi_t(\hat{W}_\tau)$ i dependent not only of the wage profile but al of the date τ at which the match was form Moreover, since the absence of external re utation effects implies that the market alt natives $\bar{U}_t$ and $\bar{\Pi}_t$ at each t are the same for matches no matter when they were formed, follows that

$$(\text{A1}) \qquad U_t(\hat{W}_t) + \Pi_t(\hat{W}_t)$$
$$= U_t(\hat{W}_\tau) + \Pi_t(\hat{W}_\tau),$$
$$\text{for all } \tau \leq$$

Together with the incentive compatibility co dition (9), this implies

$$(\text{A2}) \qquad U_t(\hat{W}_t) + \Pi_t(\hat{W}_t) - \bar{U}_t - \bar{\Pi}_t \geq \frac{v}{\delta\rho_t}$$
$$\text{for } t >$$

which establishes that there is a rent of $v/$ to forming new matches at $t > 1$.

Consider a market equilibrium with $J_t <$ As already shown, this has $\bar{\Pi}_t = \Pi_t(\hat{W}_t)$. B then (A2) implies $U_t(\hat{W}_t) \geq \bar{U}_t + v/\delta\rho_t$ $t > 1$. Thus workers necessarily receive hig utility from getting a job than from being employed even at the start of a match. Co versely, a market equilibrium with $J_t > L$ $\bar{U}_t = U_t(\hat{W}_t)$ and use of this in (A2) es implies that $\Pi_t(\hat{W}_t) \geq \bar{\Pi}_t + v/\delta\rho_t$ for $t > 1$. Thus firms necessarily receive higher profits from filling a job than from leaving it vacant even at the start of a match. To summarize, a market equilibrium with more jobs than workers has all workers employed and firms receiving a rent of at least $v/\delta\rho_t$ from filling a vacant job at any $t > 1$. In contrast, a market equilibrium with more workers than jobs has all jobs filled and workers receiving a rent of at least $v/\delta\rho_t$ from becoming employed at any $t > 1$.

---

[19] There is also a growing literature that studies self-enforcing agreements in the context of social exchange, for example, Crémer (1986), Kandori (1992b), Avner Greif (1994), and Rachel E. Kranton (1996).

This argument applies only to matches formed in periods $t > 1$, that is, to periods after the first period in which the market operates. It does not apply to matches formed in period 1 because decisions to shirk and cheat depend only on future payoffs and so are unaffected by payoffs in period 1. [This is just the point made in Carmichael (1985) about bonding.] But, in any ongoing market, $t = 1$ is a date in the past and, as time goes by, matches formed at that date become increasingly irrelevant to current market conditions because of exogenous turnover.

REFERENCES

Abowd, John M.; Kramarz, Francis and Margolis, David N. "High Wage Workers and High Wage Firms." Mimeo, Cornell University, 1993.

Abreu, Dilip. "On the Theory of Infinitely Repeated Games with Discounting." *Econometrica*, March 1988, *56*(2), pp. 383–96.

Akerlof, George A. "Labor Contracts as Partial Gift Exchange." *Quarterly Journal of Economics*, November 1982, *97*(4), pp. 543–69.

Bewley, Truman F. "A Depressed Labor Market as Explained by Participants." *American Economic Review*, May 1995 (*Papers and Proceedings*), *85*(2), pp. 250–54.

———. "A Depressed Labor Market As Explained by Participants." Mimeo, Yale University, 1997.

Blinder, Alan S. and Choi, Don H. "A Shred of Evidence on Theories of Wage Stickiness." *Quarterly Journal of Economics*, November 1990, *105*(4), pp. 1003–15.

Boyd, Robert and Richerson, Peter. *Culture and the evolutionary process*. Chicago: University of Chicago Press, 1985.

Brown, Charles. "Firms' Choice of Method of Pay." *Industrial and Labor Relations Review*, February 1990, *43*(3), pp. 165S–82S.

———. "Wage Levels and Methods of Pay." *Rand Journal of Economics*, Autumn 1992, *23*(3), pp. 366–75.

Bull, Clive. "The Existence of Self-Enforcing Implicit Contracts." *Quarterly Journal of Economics*, February 1987, *102*(1), pp. 147–59.

Bulow, Jeremy I. and Summers, Lawrence H. "A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynesian Unemployment." *Journal of Labor Economics*, July 1986, *4*(3), pp. 376–414.

Campbell, Carl M., III and Kamlani, Kunal S. "The Reasons for Wage Rigidity: Evidence from a Survey of Firms." *Quarterly Journal of Economics*, August 1997, *112*(3), pp. 759–89.

Card, David and Krueger, Alan B. *Myth and measurement: The new economics of the minimum wage*. Princeton, NJ: Princeton University Press, 1995.

Carmichael, Lorne. "Can Unemployment Be Involuntary? Comment." *American Economic Review*, December 1985, *75*(5), pp. 1213–14.

———. "Efficiency Wage Models of Unemployment—One View." *Economic Inquiry*, April 1990, *28*(2), pp. 269–95.

Crémer, Jacques. "Cooperation in Ongoing Organizations." *Quarterly Journal of Economics*, February 1986, *101*(1), pp. 33–49.

Davis, Steve J. and Haltiwanger, John. "Wage Dispersion between and within U.S. Manufacturing Plants, 1963–86." *Brookings Papers on Economic Activity*, Microeconomics 1991, pp. 115–80.

Dickens, William T. and Katz, Lawrence F. "Inter-Industry Wage Differences and Industry Characteristics," in Kevin Lang and Jonathan S. Leonard, eds., *Unemployment and the structure of labor markets*. New York: Blackwell, 1987, pp. 48–89.

Fehr, Ernst; Gächter, Simon and Kirchsteiger, Georg. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, July 1997, *65*(4), pp. 833–60.

Greif, Avner. "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies." *Journal of Political Economy*, October 1994, *102*(5), pp. 912–50.

Grossman, Sanford J. and Hart, Oliver D. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *Journal of Political Economy*, August 1986, *94*(4), pp. 691–719.

Harris, John R. and Todaro, Michael P. "Migration, Unemployment and Development: A Two-Sector Analysis." *American Economic Review*, March 1970, *60*(1), pp. 126–42.

Hayek, F. A. *Law, legislation and liberty*. London: Routledge and Kegan Paul, 1982.

Kahneman, Daniel; Knetsch, Jack L. and Thaler, Richard. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review*, September 1986, *76*(4), pp. 728–41.

Kandori, Michihiro. "Repeated Games Played by Overlapping Generations of Players." *Review of Economic Studies*, January 1992a, *59*(1), pp. 81–92.

———. "Social Norms and Community Enforcement." *Review of Economic Studies*, January 1992b, *59*(1), pp. 63–80.

Kaufman, Roger T. "On Wage Stickiness in Britain's Competitive Sector." *British Journal of Industrial Relations*, March 1984, *22*(1), pp. 101–12.

Kranton, Rachel E. "Reciprocal Exchange: A Self-Sustaining System." *American Economic Review*, September 1996, *86*(4), pp. 830–51.

Kremer, Michael. "The O-Ring Theory of Economic Development." *Quarterly Journal of Economics*, August 1993, *108*(3), pp. 551–75.

Kremer, Michael and Maskin, Eric. "Wage Inequality and Segregation by Skill." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5718, August 1996.

Krueger, Alan B. and Summers, Lawrence H. "Efficiency Wages and the Inter-Industry Wage Structure." *Econometrica*, March 1988, *56*(2), pp. 259–93.

Kuznets, Simon. "Economic Growth and Income Inequality." *American Economic Review*, March 1955, *45*(1), pp. 1–28.

Lang, Kevin. "Persistent Wage Dispersio
Involuntary Unemployment." *Quar*
Journal of Economics, February 19
106(1), pp. 181–202.

MacLeod, W. Bentley. "Complexity, C
tract and the Employment Relationshi
Discussion Paper No. 97-22, The L
School, University of Southern Califo
1997.

MacLeod, W. Bentley and Malcomson, James
"Implicit Contracts, Incentive Compati
ity, and Involuntary Unemploym
Econometrica, March 1989, *57*(2), p
447–80.

———. "Motivation, Markets, and D
Economies." University of Southamp
Discussion Papers in Economics a
Econometrics No. 9319, August 1993a.

———. "Wage Premiums and Profit Ma
mization in Efficiency Wage Models." *E*
ropean Economic Review, August 199
37(6), pp. 1223–49.

MacLeod, W. Bentley and Parent, Daniel. "J
Characteristics and the Form of Comp
sation." Mimeo, University of South
California, 1997.

MacNeil, Ian R. "The Many Futures of Co
tracts." *Southern California Law Revie*
1974, *47*(688), pp. 691–816.

Millward, Neil; Stevens, Mark; Smart, Da
and Hawes, W. R. *Workplace industrial r*
lations in transition: The ED/ESRC/PS
ACAS surveys. Aldershot, U.K.: Dar
mouth, 1992.

Moene, Karl Ove. "A Reformulation of th
Harris-Todaro Mechanism with Endog
nous Wages." *Economics Letters*, 198
27(4), pp. 387–90.

Montgomery, James D. "Equilibrium Wa
Dispersion and Interindustry Wage Diff
entials," *Quarterly Journal of Econom*
February 1991, *106*(1), pp. 163–79.

Pearce, David G. "Repeated Games: Coop
tion and Rationality," in Jean-Jacq
Laffont, ed., *Advances in economic theo*
Sixth World Congress, Vol. I. Cambrid
Cambridge University Press, 1992, pp. 13
74.

Raff, Daniel M. G. "Wage Determination Th
ory and the Five-Dollar Day at Ford." *Jo*
nal of Economic History, June 1988, 48(
pp. 387–99.

Daniel M. G. and Summers, Lawrence H. "Did Henry Ford Pay Efficiency Wages?" *Journal of Labor Economics*, October 1987, *5*(4), pp. S57–S86.

ey, Garey and Watson, Joel. "Contractual Fragility, Job Destruction, and Business Cycles." *Quarterly Journal of Economics*, August 1997, *112*(3), pp. 873–911.

Roth, Alvin E. "Introduction to Experimental Economics," in John H. Kagel and Alvin E. Roth, eds., *Handbook of experimental economics*. Princeton, NJ: Princeton University Press, 1995, pp. 3–109.

Shapiro, Carl and Stiglitz, Joseph E. "Equilibrium Unemployment as a Worker Discipline Device." *American Economic Review*, June 1984, *74*(3), pp. 433–44.

Stewart, James B. "Taking the Dare." *New Yorker*, July 26, 1993, pp. 34–39.

Troske, Kenneth R. "Evidence on the Employer Size-Wage Premium from Worker-Establishment Matched Data." U.S. Bureau of the Census (Washington, DC) Technical Report, April 1994.

Weitzman, Martin L. "A Theory of Wage Dispersion and Job Market Segmentation." *Quarterly Journal of Economics*, February 1989, *104*(1), pp. 121–37.

Williamson, Oliver E. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press, 1975.

The Economics of Child Labor

By KAUSHIK BASU AND PHAM HOANG VAN*

If child labor as a mass phenomenon occurs not because of parental selfishness but because of the parents' concern for the household's survival, the popular argument for banning child labor loses much of its force. However, this assumption about parental decision-making coupled with the assumption of substitutability in production between child and adult labor could result in multiple equilibria in the labor market, with one equilibrium where children work and another where adult wage is high and children do not work. The paper establishes this result and discusses its policy implications. (*JEL* J20, K31, D60)

According to the International Labour Organization (ILO), in 1990 there were almost 79 million children around the world who did regular work (see Kebebew Ashagrie, 1993 p. 16). This estimate of child labor would vary depending on how we define work, how we define a child, and how we collect the data, but no matter which estimate we take, the inescapable fact remains that this is a problem of gigantic proportions. Moreover, the magnitude of the tragedy is not captured by numbers alone, since the conditions of child labor can vary. There are children who work in hazardous industries, risking accident and injury; there are others working in conditions that take a slower but definite toll on the children's health.

As people become informed about child labor, the natural reaction is to seek ways to banish child labor. The easiest way to banish
or so it seems—is to ban it. And across
world there has been an increasing choru
opinion seeking a ban on child labor. De
of the proposals have varied. In the Uni
States, the so-called Harkin's bill (Child L
Deterrence Act of 1997) seeks to ban the
port of those goods which have used child
bor as input. International organizations
many citizens fora have talked about label
products which are free from child-labor
puts so that individuals, by confining their
sumption to such goods, can bring about
effective ban. It will be argued later that m
of these well-meaning interventions can
counterproductive.

This is a field of study where prescrip
has outstripped analysis by a wide margi
is the aim of this paper to construct a mo
of child labor which can then be used to
and answer some policy questions. Ther
one central idea which is at the heart of
model. The next two paragraphs give an
tuitive sketch of this idea.

In the popular mind, child labor is very
equated with child abuse. The phenomeno
taken to be a product of avaricious ent
neurs seeking cheap labor and selfish pa
who would prefer enjoying leisure while
children work. It seems to us that while
popular description of entrepreneurs may
be accurate, the parents are mischaracte
We argue instead that the traditional mod
the household, where parents take their
dren's interests into account, while some

---

* Department of Economics, Cornell University, Ithaca, NY 14853. This paper has benefited from seminar presentations at Boston University, Cornell University, the Delhi School of Economics, DELTA, Georgetown University, Pennsylvania State University, the University of California-Berkeley, the University of Namur, and Yale University. We would like to thank Dilip Abreu, Jim Albrecht, Pranab Bardhan, Valerie Bencivenga, François Bourguignon, Martin Browning, Veena Das, Rajat Deb, Ron Ehrenberg, James Foster, Patrick Francois, Albert Hirschman, Alain de Janvry, Heraklis Polemarchakis, Debraj Ray, T. N. Srinivasan, Erik Thorbecke, John Toye, Susan Vroman, and Henry Wan, Jr. for comments and criticisms at different stages of this work. We are also grateful to three anonymous referees of the journal for helpful suggestions.

realistic, is a better model. Thus, while not denying that child abuse does occur in all societies, we take the position that when we have children working as a mass phenomenon as in many less-developed countries, it is much more likely that this reflects not a difference in the attitude of the parents but the problem of stark poverty where the parents are compelled to send the children to work for reasons of survival. Even in England, which witnessed some of the worst excesses of child labor in the late eighteenth and early nineteenth centuries, a parliamentary report noted that "parents were desperately unhappy about the situations their children were in but could do nothing about it. The social system allowed them no choice." (Sara Horrell and Jane Humphries, 1995.)

Once we accept this description of household decision-making, the case for declaring child labor illegal gets considerably weakened but in some situations there may nevertheless be a more complicated and equilibrium-based reason for declaring child labor illegal. If we agree that sending children out to work is an act of desperation on the part of the parents, it seems reasonable to expect that parents would not send their children to work if their own wages were higher or employment prospects better. Now do the following experiment. Suppose all children are pulled out from work, say because of a total ban. What effect will this have? Clearly, the first effect of this will be a shortage of labor. And given that child and adult labor are usually substitutes, the wages of adults will rise in response to the excess demand for labor.[1] But as adult wages rise, it is possible, given our above assumption, that parents will not now want to send their children to work. Hence, the ban may become redundant. In brief, once a ban is imposed, the ban may become unnecessary. Essentially what we are claiming is that the labor market may be characterized by multiple equilibria—one in which wages are low and children work and another in which wages are high and children do not work.

In the scenario described here, the purpose of government intervention is very different from that in conventional models. In our model, intervention does not create a new equilibrium but simply jolts the economy out of one equilibrium to another *preexisting* equilibrium. In this model, partial bans can have unexpected adverse effects.

Economists seeking government intervention in the child-labor market have typically justified their recommendation by claiming that there are externalities to child labor or that private returns to education are smaller than social returns.[2] But such arguments need to be substantiated since "externalities" are too often treated as a catchall. What our model demonstrates is that in certain specific situations there may be a rigorous case for a ban simply based on the child-labor market's natural tendency to exhibit multiple equilibria.

There are many other aspects of child labor which are important—its dynamic implications, its relation to education and human capital, and its medical aspects. But those are not our concern here. Our focus is on the multiple equilibria which seems to be a natural and inherent (potential) characteristic of child-labor markets but have eluded researchers and observers in this field.

The plan of the paper is as follows. Section I presents some basic information on child labor and some accounts of historical experience which are relevant as backdrop to our model. Section II presents a basic model and introduces a diagrammatic technique for depicting equilibria. Section III suggests ways of generalizing the basic model. Policy questions and the subject of legislation form the subject matter of Section IV. Section V considers the implications of the model for the economics of fertility and suggests ways of extending this kind of modeling to other areas.

I. Facts and Experience

To begin with the current scenario, the only thing that one can be certain about are the broad parameters of the problem. We know that a very large number of children—meaning persons below the age of 15 years—work. Most of these working children are in the Third World, with the exception of child prostitution, the incidence of which can be high even in industrialized nations. The bulk of child laborers belong to the 10-to-14-year age category; but there is also a substantial number of children below 10 years of age who work.

---

[1] In case wages are rigid we would expect adult unemployment to decline.

[2] A more sophisticated claim is that child labor is a manifestation of failures in other markets, such as the market for capital or insurance (Christiaan Grootaert and Ravi Kanbur, 1995).

As we go behind such broad generalizations to actually construct numbers, we run into controversy. Employment surveys typically do not have respondents below 15 years of age. Some countries, such as India, have tried to officially count the number of children who work. But one can get very different answers depending on which source one turns to (for discussion, see Myron Weiner, 1991; Grootaert and Kanbur, 1995). For instance, in 1983 the national sample survey estimates showed that 17.4 million children worked, whereas a study by the Operations Research Group, conducted at the behest of the Ministry of Labour, estimated the number to be close to 44 million.

For an overall statistical picture, one can turn to the ILO estimates of 1993 collated and quoted in Ashagrie (1993 Table 4). Among children between 10 and 14 years of age, 70.9 million are laborers. If we look at "participation rates," that is, the *percentage* of children who work among all children of that age-group, the figures can be quite alarming. For the world as a whole for the 10-to-14-year age-group, the participation rate is 13.7 percent, and in some parts of central Africa the figure can be as high as 32.9 percent.[3]

Historically, child labor was not the preserve of Africa, Latin America, and Asia. Some of the worst excesses occurred in Europe in the late eighteenth and early nineteenth centuries and especially in Britain during the Industrial Revolution.[4] According to most sources, the participation rates in Britain d ing its industrial revolution were very high higher than the contemporary rates in regions of the world with the sole exceptio middle Africa. According to the 1851 cen in England and Wales 36.6 percent of b aged 10–14 and 19.9 percent of girls in same age-group were working. It is striking note that these high participation rates in 18 existed despite the main Factories Acts (1833 and 1844), which placed curbs on ch labor, being already in place, and child la arguably being on the wane.[5]

One important question is: what affect d the Factories Acts have on the incidence child labor? The answer to this will help speculate about the consequences of the ma laws which are currently either in effect or der consideration. A study by Grootaert Kanbur (1995) suggests that the incidence child labor was declining even before the F tories Acts. Given that the nonpoor people poor countries do not send their children work, could we assert that child labor in Br ain would vanish anyway as British prospe rose, with or without laws to curtail childre employment? The model we develop sho help us ponder such questions, but in the r mainder of this section let us try to elicit formation from the historical literature in or to give shape to some of the *assumptions* th we use to build our model.

The popular instinct among most sections our society is to support ideas such as th outlined in Senator Harkin's bill in the Un States, which seeks to ban the import of ch labor-tainted products. This popular inst stems from the presupposition that the ex tence of child labor is the product of greed the part of employers who employ the child and the parents who send the children to wo As stated in the introduction, we reject view of the parents. And indeed there is overwhelming support for this rejection.

---

[3] For a survey of the contemporary world situation pertaining to child labor, which goes beyond numbers and looks at institutional details, see Assefa Bequele and Jo Boyden (1988).

[4] We confine most of our historical observations to Britain. The reader may refer to Weiner (1991) for a brief description of the experience of other European nations and also Japan and the United States.

[5] A districtwise breakdown of this data is repo Hugh Cunningham (1990). Though we say that chi bor was on the decline by 1851, it is possible th number of children who did some work peaked in (see Clark Nardinelli, 1990). However, the mitigating tor was that, by the late nineteenth century, most ch were working only halftime. This was in response requirement of the Factories Act of 1874 that childr tend school on at least a half-time basis.

The first and foremost evidence is the contemporary fact that the children of the nonpoor seldom work even in very poor countries. This phenomenon is best explained by supposing that parents withdraw their children from the labor force as soon as they can afford to do so. In other words, children's leisure or, more precisely, nonwork[6] is a luxury good in the household's consumption in the sense that a poor household cannot afford to consume this good but it does so as soon as the household income rises sufficiently. In our second model, we use the Stone-Geary utility function to capture this idea.

Another source of evidence comes from the late nineteenth-century census data for Philadelphia. Claudia Goldin's (1979) analysis of this data leads her to conclude (p. 124): "The higher the father's wage, the lower the probability of the child participating in the labor force."; and also: "The father's unemployment sent both boys and girls into the labor force, with a stronger impact on the former." A different kind of evidence comes from David Vincent's (1981) study of working-class autobiographies. The study showed that children who worked rarely blamed their parents, believing instead that it was poverty that drove the parents into making the children labor (see also Michael Anderson, 1971).[7]

By attributing to each household one utility function, our analysis does abstract from reality. There is evidence, for instance, that household consumption patterns differ depending on who takes the decisions and who earns the money.[8] Despite this abstraction, it is worth emphasizing that our model does not conflict with recent evidence and theories which ask for the rejection of the "unitary model" of the household. This is because we assume that a child's labor-supply decision is taken by a parent. There is no attempt to deny that this decision could be different if the decision-making were shifted to another member of the household.

More generally, all we want is to give primacy to the household or family wealth as a determinant of child labor. There has been some recent attempt to model parents and the children as being involved in bargaining conflicts (Carolyn M. Moehling, 1995; Manash R. Gupta, 1998). Such investigations are worthwhile but, if we were to have one representative model for analyzing child labor, we do not consider the bargaining model to be the right one. The model presented in the next section captures the essentials of our main theoretical idea.

Finally, it is important to emphasize that the phenomenon of child labor has important sociological and psychological issues at stake. The child-labor market does not always operate on the basis of voluntary exchange but involves coercion and psychological pressures (see Jonathan Silvers, 1996 p. 82). Nevertheless, we have stayed away from many of the larger issues and confine our attention to a rigorous, economic analysis because it is not clear to us how we can take on board different aspects of this important phenomenon—economic, sociological, psychological—all at once. There is no choice but to dissect such a large phenomenon into several parts and to analyze these one at a time. Moreover, we hope that our paper demonstrates how well-meaning spontaneous recommendations can often backfire. This is an area where what seems obviously the right thing to do may turn out, on deliberation, to be quite the opposite. As a consequence, this is also an area where individuals and groups, with their own self-interested agenda, can garner mass support for policies which actually benefit them while superficially appearing to help the cause of the laboring children. Formalism and scientific inquiry can be a bulwark against this.

---

[6] Since the alternative to work may not necessarily be leisure. It could, for instance, be education.

[7] The only exception to these findings occurs in the case of alcoholism on the part of parents. It is difficult to get data on alcoholism. We have simply been able to determine that in 1800, an average person in England and Wales consumed 27 gallons of beer per annum (Nardinelli, 1990). But it is difficult to judge from this alone (without information on the distribution of this consumption and the consumption of other types of alcohol) as to how heavy the drinking was. However, the sociological and historical writings cited above do not give the impression of alcoholism being particularly high and, therefore, the cause of mass child labor.

[8] For discussion on this see, e.g., Amartya K. Sen (1990), Martin Browning et al. (1994), and Christopher Udry (1996).

II. Child Labor: A Basic Model

What is nice about the results derived from this model and the one in the next section is that they are based on very weak assumptions. The two essential assumptions may be codified as the following two axioms.

The Luxury Axiom: A family will send the children to the labor market only if the family's income from non-child-labor sources drops very low.

The Substitution Axiom: From a firm's point of view, adult labor and child labor are substitutes. More specifically, child labor can be substituted by adult labor.

In constructing the models we shall use many special assumptions and functional forms but those are all expositional devices. They keep the analysis tractable. All our main results are, we believe, essentially derived from the luxury and substitution axioms. It is worth stressing here that the luxury axiom that we need is weaker than the word "luxury" suggests. (This is clarified in footnote 11.) These assumptions are not in themselves sufficient for generating multiple equilibria but they are sufficient for giving us a model with a *potential* multiplicity of equilibria. We discuss the conditions under which multiple equilibria actually occur, after describing the model formally. The above assumptions are built into the preference relations and production functions that we specify in this and the next section.

Assume that there are N identical families (or households) in the economy and that each family consists of one adult and one child. The latter of course may be simply a convention whereby *we call* the two parents "one adult" and the two children "one child." The family's preference, $\succ$, is described by a binary relation defined on the set

$$(1)\qquad \{(c, e) \mid c \geq 0, e \in \{0, 1\}\},$$

where c is consumption by each family member and e is the child's work effort which can only take on values of 0 or 1. We are assum that the adults always work, no matter the wages are. And for simplicity, child adult consumptions are presently assumed be equal.

We shall now impose an assumpti which is in keeping with the luxury axi and the arguments presented in this sectio It is, however, very strong. This is only reasons of simplicity and is relaxed late The assumption is as follows. A family p fers to send the child to work if and only in the absence of income from the chil each individual's consumption falls belo a certain exogenously fixed subsisten level, s. More formally, for all $\delta > 0$,

$$(2)\qquad (c, 0) \succ (c + \delta, 1) \qquad \text{if } c \geq s,$$

$$\text{and } (c + \delta, 1) \succ (c, 0) \qquad \text{if } c <$$

The household's aim is to choose c and e as to maximize its preference subject to t following budget constraint:

$$(3)\qquad 2c \leq ew_C + w_A,$$

where w_C and w_A are the market wages f respectively, child and adult labor. Ea household treats these wages as given.

The solution to the household's maximiz tion problem, therefore, is as follows:

$$(4)\qquad c(w_A) = \begin{cases} \dfrac{w_A}{2} & \text{if } w_A \geq 2s \\ \dfrac{w_A + w_C}{2} & \text{if } w_A < 2s \end{cases}$$

$$(5)\qquad e(w_A) = \begin{cases} 0 & \text{if } w_A \geq 2s \\ 1 & \text{if } w_A < 2s. \end{cases}$$

It follows that labor supply of adults children, denoted by S^A and S^C, are given

$$(6)\qquad S^A = N;$$

$$(7)\qquad S^C(w_A) = \begin{cases} 0, & \text{if } w_A \geq 2s \\ N, & \text{if } w_A < 2s. \end{cases}$$

Our next step is to derive the market demand for adult and child labor. To do so we invoke the substitution axiom and make the simplifying assumption that adults and children are substitutes in production subject to an adult-equivalent scaling, given by γ, where $0 < \gamma < 1$. So assume there are n identical firms, each producing a single consumption good. Each firm i's production function is given by:

$$(8) \quad x_i = f(A_i + \gamma C_i), f' > 0, f'' < 0,$$

where x_i is firm i's output of the consumption good, and A_i and C_i are respectively the numbers of adult and child laborers employed by firm i. The firm is a wage taker. Hence, firm i's problem is as follows:

$$(9) \quad \max_{\{A_i, C_i\}} f(A_i + \gamma C_i) - A_i w_A - C_i w_C.$$

The solution to (9) is straightforward. If $w_A < w_C/\gamma$, then the firm will employ only adults. If $w_A > w_C/\gamma$, then it will employ only children. If $w_A = w_C/\gamma$, then it will be indifferent between adults and children. We call w_C/γ the "effective child wage," that is, the market child wage per adult-equivalent. In addition, each firm will always ensure that

$$(10) \quad f'(A_i + \gamma C_i) = \min\left\{w_A, \frac{w_C}{\gamma}\right\}.$$

The aggregate demand for adult and child labor, D^C and D^A, is derived by multiplying each firm's demand by n. Hence, $D^A = D^A(w_A, w_C)$ and $D^C = D^C(w_A, w_C)$ are given implicitly by the following.

$$(11) \quad \text{If } w_A > \frac{w_C}{\gamma} \quad \text{then } D^A = 0$$

$$\text{and } f'\left(\frac{\gamma D^C}{n}\right) = \frac{w_C}{\gamma}.$$

$$\text{If } w_A < \frac{w_C}{\gamma} \quad \text{then } D^C = 0$$

$$\text{and } f'\left(\frac{D^A}{n}\right) = w_A.$$

$$\text{If } w_A = \frac{w_C}{\gamma}$$

$$\text{then } f'\left(\frac{D^A + \gamma D^C}{n}\right) = w_A = \frac{w_C}{\gamma}.$$

A *labor-market equilibrium* in this simple model is a pair of wages, (w_A^*, w_C^*), such that

$$(12) \quad D^A(w_A^*, w_C^*) = N, \quad \text{and}$$

$$D^C(w_A^*, w_C^*) = S^C(w_A^*).$$

At first sight it may seem that what we have described is a *partial* equilibrium. However, it is easy to embed this model in a general equilibrium framework without having to modify the above description. One way is to think of this as an economy where the firms' profits are not shared with the households but instead are consumed entirely by the entrepreneurs of the firms. In that case the labor-market equilibrium would fully characterize the closed-economy general equilibrium.[9] Alternatively, we could assume this to be a small open economy which would imply that the goods market will trivially clear and the same results would derive. One implication of viewing this as a general equilibrium will be that both equilibria will be Pareto efficient by the fundamental theorem of welfare economics, though, of course, the labor households may be better off in one equilibrium rather than another.

We now develop a diagrammatic technique for depicting this equilibrium. The geometry, apart from aiding intuition, turns out to be a very useful instrument for doing policy analysis. It also helps us see very clearly how this model *may* exhibit multiple equilibria so that in the same economy, children working and children not working can be part of equilibrium behavior.

In Figure 1, consider first the (w_A, w_C)-space. The axes of this space are marked **O**w_C and **O**w_A. For wage pairs above the horizontal line $w_A = 2s$, children will not work, $e = 0$;

---

[9] We are grateful to Heraklis Polemarchakis for discussion on this.

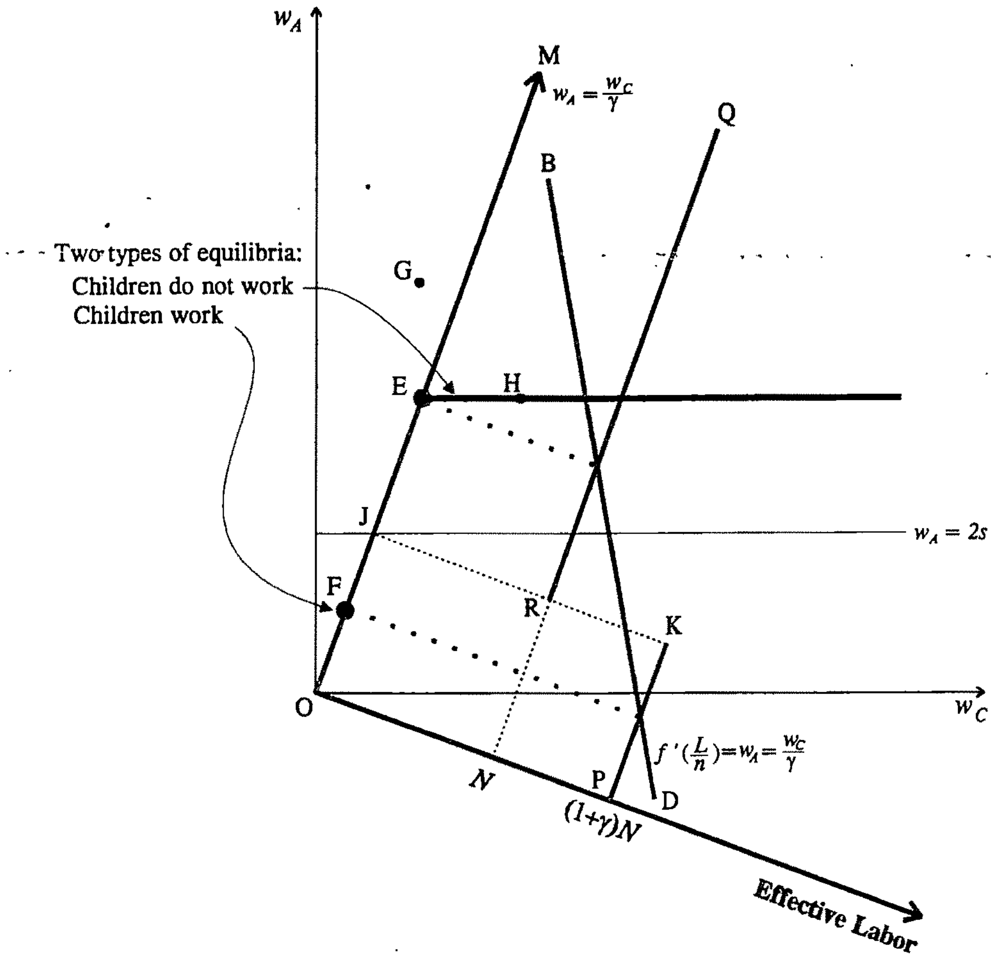


FIGURE 1. EQUILIBRIUM IN THE CHILD- AND ADULT-LABOR MARKETS

and below this line $e = 1$. In this space draw the graph of the function $w_A = w_C/\gamma$. This is the line **OM**. Since $\gamma < 1$, this line is steeper than 45°. This is a very significant line and will be referred to here as the "*ridge.*" If (w_A, w_C) is above the ridge, then the demand for adult labor is zero; if (w_A, w_C) is below the ridge, then the demand for child labor is zero. What happens if the market wages lie on the ridge? Let us define the "*effective labor*" used or demanded by a firm to be the total amount of labor measured in adult-equivalents being used or demanded by the firm. So if a firm i employs A_i adult laborers and C_i child laborers, its *effective* labor employment is $A_i + \gamma C_i$. If the market wages for adult and child laborers lie on the ridge then we know from (11) that each firm's effective demand for labor, L_i, is given by $f'(L_i) = w_A = w_C/\gamma$. In other words, firm i facing such a (w_A, w_C) will be willing to employ any combination of labor, A_i, and child labor, C_i, as long as $\gamma C_i = L_i$. Since all n firms are identical aggregate effective labor demand in the ket is $L = nL_i$, which allows L to be implic defined by $f'(L/n) = w_A = w_C/\gamma$.

Consider the ridge as the "vertical" and draw a line through **O** which is orthog to the ridge and going eastward. The thick labeled "Effective Labor" in Figure 1, resents this line. Now, to start with, con only wages which lie on the ridge. We call the two-dimensional Euclidean spa which the "vertical" axis is the ridge and "horizontal" axis effective labor the " *Euclidean space.*" For every point on ridge (showing a particular wage pair) off the firms' effective labor demand on axis marked "Effective Labor." That give us a downward-sloping curve in the

Euclidean space. The line **BD** is an *n*-fold "horizontal" ("horizontal" within quotes will from now on always mean horizontal in the tilted Euclidean space) blowup of such a line and therefore represents the firms' aggregate effective demand for labor for wage pairs lying on the ridge. Hence, if (w_A, w_C) is point **E**, the aggregate effective demand for labor is given by **ON**.

It is now easy to read off the respective demand for labor for wage pairs which are not on the ridge. In the (w_A, w_C)-space, suppose (w_A, w_C) happens to be a point vertically above **E**. Then clearly, adult wage exceeds effective child wage, $w_A > w_C/\gamma$. Hence, from (11) and (12), the firms' aggregate demand for adult labor is zero and the firms' aggregate demand for child labor expressed in adult-equivalents is **ON**. Thus the wage determining the amount of effective labor demanded by firms is the child wage. Or, to put it differently, given a wage pair at **G**, the aggregate effective demand for labor is given by moving vertically down from **G** to **E** on the ridge and then "horizontally" to the line **BD**. Hence, the effective demand is **ON**. For points above the ridge the effective demand is exclusively effective demand for *child* labor. The reader should satisfy himself or herself that for wage pairs below the ridge the same exercise is carried out by moving horizontally to the ridge and, of course, the demand for labor is now exclusively for adult labor. Hence, given a wage pair at **H**, the demand for child labor is zero and the demand for adult labor is **ON**.

In Figure 1, let us now draw the effective labor supply (that is, aggregate labor supply measured in adult-equivalents) corresponding to wage pairs that lie on the ridge. Note that for all wage pairs on the ridge and above **J**, the supply of child labor is zero [see (7)]. Hence, for all such wage pairs the effective supply of labor is **ON**, where $\overline{\mathbf{ON}} = N$ (that is, the number of adults in the economy). If the wage pair is below **J**, the aggregate effective supply of labor is given by $N + \gamma N$ since all children are now out to work. Hence, the aggregate, effective supply of labor in the tilted Euclidean space is given by the two line segments **QR** and **KP**.

We shall first locate equilibria that may lie on the ridge. This is done simply by looking at the tilted Euclidean space and the points of intersection between the aggregate (effective) demand and supply curves. In the case illustrated in Figure 1 there are two equilibria given by the wage pairs **E** and **F**. At **F** both adults and children work, adult wage is very low and children's wage even less. At **E**, adult wage is high, no children offer labor on the labor market, and the entire demand for labor is met by the supply of adults.

To complete the search for equilibria, we must now check if there are any equilibria off the ridge. Using the "ridge equilibria" as benchmark, this is easy to do. All wage pairs on the horizontal line through **E** and **H** and to the right of **E** constitute equilibrium wage pairs. Since, in this simple model, these are trivial extensions of the equilibrium at **E**, we shall in the remainder of this section focus attention only on the "ridge equilibria."

The occurrence of multiple equilibria is by no means necessary in this model. If a country's labor force becomes more productive (because of better technology, for instance), so that the aggregate demand curve, **BD**, shifts to the "right" (that is, in the tilted Euclidean space), we shall soon have an economy with a unique equilibrium where only adults work. We believe that industrialized countries are in such a situation. If on the other hand, labor is very unproductive, so **BD** shifts to the "left," we could have a unique equilibrium and child labor is a necessary phenomenon.[10]

There may not be a case for banning child labor in such a situation. As can be checked from Figure 1, a ban in such a model will raise adult wage but will nevertheless be less than $2s$. As long as this new adult wage is less than the previous adult wage plus the child wage, all laboring households will be worse off. The popular support for a child labor ban in such situations usually stem from other hidden agenda such as protectionism or misguided concern for labor. Any argument for a ban has to be much more sophisticated. We discuss this matter in detail in Section IV.

---

[10] It is arguable that Britain in the early nineteenth century had only the bad equilibrium; then in the mid-nineteenth century the bad and good equilibria; and by the start of the twentieth century only the good one. Policy intervention would be important mainly in the middle case. It would be redundant by the century's end, and very difficult to effectively implement and also of dubious welfare consequences at the start of the nineteenth century.

Return now to the case of two equilibria as shown in Figure 1, and suppose that the economy is currently at the "lower" equilibrium—that is, at **F**. While a model is never an exact mirror of reality, it is possible that Europe towards the end of the nineteenth century, in the last years of its industrial revolution, resembled this equilibrium better than any other. Wages were low; children worked for wages but labor productivity was moderately high.

The policy issue here is very interesting. A ban on child labor can very well be justified. If there is a total ban on child labor, effectively, the supply curve of labor in Figure 1 will be the "vertical" line from **Q** through **R**, all the way down to **N**. Hence, the equilibrium at F ceases to be an equilibrium. The only possible equilibrium occurs at **E**. At this equilibrium there is no child labor. What is interesting, however, is that the legislation banning child labor, which has so big an effect moving the economy from **F** to **E**, ceases to be a legislation of any consequence after this change. That is, even if the legislation were to be subsequently revoked, the economy would remain at **E**. This is a consequence of multiple equilibria. "Interventionist policy" clearly acquires a new meaning in economies with multiple equilibria. Such a policy will be called *benign* intervention, since such a policy ceases to constrain anybody's behavior simply by virtue of being there. Its entire effect is in terms of its *initial* impact. We return to further discussion of policy and welfare in a later section.

At the cost of more algebra, several aspects of this model can be generalized. A model which allows for a more realistic utility function and larger family size is developed in the next section. One can also raise the question of heterogeneity in family size, preferences, and productivity. To the extent that our central claim is an existential one, that is, one that asserts that there may exist multiple equilibria, it is not essential for our purpose to pursue such a generalization. Also the model in this section and the next makes it evident that our model is not critically dependent on the homogeneity of agents. However, to raise *further* questions of policy and impact on different kinds of households it will be natural to eralize along these lines in the future.

III. Child Labor: Sketch of a General Mod

In the general model each household is sumed to have one adult and m (≥ 1) child and each child consumes $\beta(<1)$, of what adult in the household consumes. Let c be adult's consumption, and $e \in [0, 1]$, child's effort. Hence, $1 - e$ is each child's sure. Effort is now chosen from a continu of possibilities. We shall represent the hou hold preference by the Stone-Geary util function:[11]

$$(13)\quad u(c, e) = \begin{cases} (c - s)(1 - e), & \text{if } c \geq \\ c - s, & \text{if } c < \end{cases}$$

where $c \geq 0$, $e \in [0, 1]$ and $s > 0$ is a par eter. The household maximizes u with resp to c and e subject to the budget constraint

$$(14)\quad c + m\beta c = mew_C + w_A.$$

From the first-order conditions we get following effort function.

$$(15)\quad e(w_A, w_C, m) = \begin{cases} 0 & \text{if } s + sm\beta + mw_C \leq w_A \\ 1 & \text{if } s + sm\beta - mw_C \geq w_A \\ \dfrac{mw_C - w_A + s + sm\beta}{2mw_C} & \text{otherwise} \end{cases}$$

---

[11] It is easy to check that this implies that the ch leisure is a luxury good because a doubling of hous wealth (from non-child-labor sources) leads to a than doubling of child leisure. However, as will be parent as we go along, we do not really need the ch leisure to be a luxury good "everywhere." Essen what we need is that there exists a positive hous wealth where children begin to consume leisure higher wealth where they cease to work.

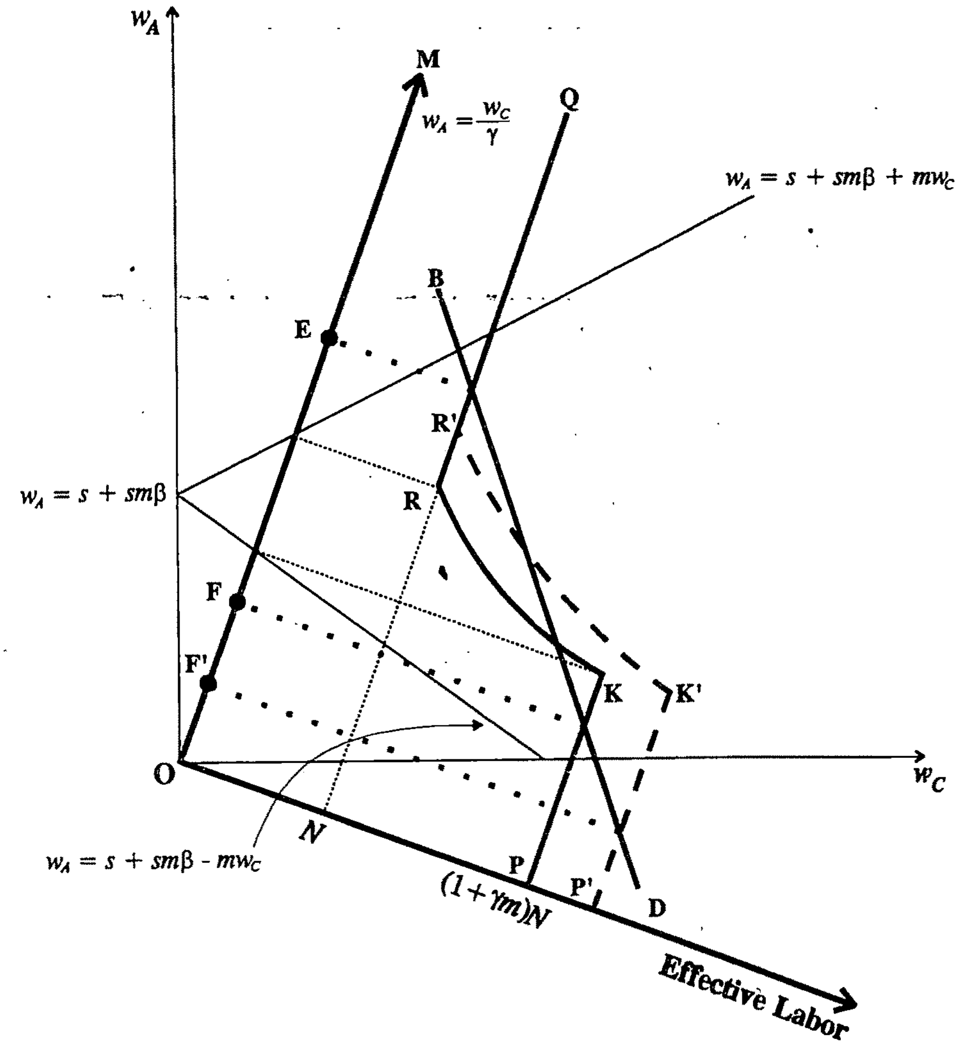

w_A
M
$w_A = \frac{w_C}{\gamma}$
Q
$w_A = s + sm\beta + mw_C$
B
E
R'
$w_A = s + sm\beta$
R
F
F'
K
K'
O
w_C
N
$w_A = s + sm\beta - mw_C$
P
P'
$(1+\gamma m)N$
D
Effective Labor

FIGURE 2. EQUILIBRIUM WITH VARIABLE EFFORT AND FAMILY SIZE

The aggregate labor-supply functions for adult and child labor are respectively:

$$S^A = N \tag{16}$$

$$S^C(w_A, w_C, m) = mNe(w_A, w_C, m).$$

The demand for adult and child labor is exactly the same as in Section II. Hence, with m held constant we define (w_A^*, w_C^*) to be an equilibrium if

$$D^A(w_A^*, w_C^*) = N, \quad \text{and} \tag{17}$$

$$D^C(w_A^*, w_C^*) = S^C(w_A^*, w_C^*, m).$$

It is worth noting here that equilibrium wages depend on m and, at times, we shall refer to the equilibrium wages as $w_A^*(m)$ and $w_C^*(m)$ to emphasize this dependence.

Using the geometric technique developed in Section II we can represent the aggregate effective labor supply and demand. Supply is given by the line **QRKP** in Figure 2. By inserting the demand curve for labor, **BD** (as before), it is clear that we shall have an odd number k of equilibria, of which $(k + 1)/2$ will be stable. The stable equilibria are denoted by points **E** and **F**.

The generalized model can be used to analyze policy as well as the effect of changing age structure of the population on child labor. To do this note that the length of **NR** in Figure 2 in terms of the adult wage at point **R** is clearly given by $s(1 + m\beta)/(1 - m\gamma)$.

First assume $1 - m\gamma > 0$. Then as m increases, **NR** becomes longer, and beyond some point an equilibrium where children do not work at all will vanish. This can, somewhat approximately, be paraphrased to say that child labor is more likely to occur in a society with relatively more children. If $1 - m\gamma \leq 0$ or $1/\gamma \leq m$, it is evident from Figure 2 that the equilibrium where children do not work does not exist anyway.

More generally, check that as m increases, the supply curve of labor, **QRKP**, moves (weakly) to the "right" to, for instance, the broken line **QR′K′P′**, where **K′P′** may be shorter than **KP**, but it may also be longer. Originally there are two equilibria at **E** and **F**. The high-wage equilibrium is still at point **E** but the low-wage equilibrium has moved from F to F′ where both adult and child wages have fallen.

It is now easy to see what happens to child labor. Suppose the "downward-sloping" demand curve for labor in the tilted Euclidean space is **BD**. It is evident that if, to start with, there was a bad equilibrium at point **F**, then as m increases, all children will continue to work (and of course there are more children now) and wages of both children and adults will be lower at the new low-wage equilibrium **F**′.

On the other hand, imagine that if, to start with, there was no bad equilibrium (that is, the demand curve went over **K**), then as m increases, a bad equilibrium can come into existence. In other words, a rise in the relative number of children can generate child labor. This seems to be consistent with the evidence (Grootaert and Kanbur, 1995).

IV. Policy Intervention and the Law

In the light of the above analysis, how should government intervene and how should legislation be used to enhance the well-being of families that are compelled to send children to work? The present section seeks answers to this question under the assumption of consumer sovereignty or, more appropriately, household sovereignty. In other words, in evaluating household welfare we assume that the household knows what is in its interest and we evaluate policies to enhance household welfare.

We are aware that our assumption does get violated in some situations. We wish to comment here on one kind of violation, brou our attention by Albert Hirschman.[12] In the argument is that certain impositio consumer sovereignty are at times desi because they may result in a genuine shi consumer preference or morality. Muc what we consider moral or immoral dep on what we are used to. We may call this " quired morality." Certain practices in f away societies or times which look obvio immoral to us may not appear so to those cieties. Likewise it is worth being aware we may have certain common practices in society which will appear shockingly imm to our descendants when they look back at late twentieth century. These acquired mo ities then influence our behavior and pre ence. Corporal punishment for children good example of such moral relativism. some societies such punishments are natur even desirable, and to others abominable. T may explain why child labor is not only erated in certain societies but considered ural and nothing to protest about. But if aversion to child labor is an acquired mora then one way to remove child labor is to and make it customary for children not work. If for instance, child labor is banned some time, then it is conceivable that our ju ment in that matter will change—so that some time, even if the law is revoked, would not *want* to send our children to w regardless of household income.

This leads also to multiple equilibria wo ing through social norms. This is an emine plausible argument. However, in what follo we work within the confines of traditional e nomics, where household preferences or j ments do not change. Our aim is to argue t *despite* this, we reach nontraditional po conclusions.

One of the central policy conclusions is important role of benign policy interventi Consider Figure 2 and suppose there is an gregate demand curve, **BD**, that cuts thro **QR** and **KP**. Then there are at least two tential equilibria. Suppose an economy caught in the bad equilibrium, that is, at p

---

[12] Personal communication to K. Basu, dated Feb 15, 1995.

Then a total ban on child labor could deflect the equilibrium all the way to the good equilibrium at point E. Hence, all working-class households would be better off. And the policy is self-liquidating in the sense that once in place it plays no role and constrains no one's behavior. This is, of course, a consequence of there being more than one equilibrium. All this we have discussed in Section II, and so we need not dwell on this anymore.

Suppose now that there is only one equilibrium, the bad one. This is because the aggregate demand curve (for labor) travels below point R and cuts the supply curve exactly once, on the segment KP. What will be the effect of a total ban on child labor?[13]

The ban will clearly cause adult wage to change from $f'((N + \gamma mN)/n)$ to $f'(N/n)$. Since $f'' < 0$, all we know is that the ban will cause adult wage to rise. The extent of the rise could vary depending on the nature of the production function, f. To see if the ban helps or hurts worker households, describe the utility levels of the household with and without a ban by, respectively, U^B and U^N. Since with no ban we have a bad equilibrium,

$$U^N = \frac{[m\gamma + 1]f'\left(\dfrac{N + \gamma mN}{n}\right)}{m\beta + 1} - s. \tag{18}$$

In the case with the ban, consumption per person is

$$c = \frac{f'\left(\dfrac{N}{n}\right)}{1 + m\beta}. \tag{19}$$

If this consumption level exceeds s, clearly the household benefits from a ban. If $c < s$,

$$U^B = \frac{f'\left(\dfrac{N}{n}\right)}{1 + m\beta} - s. \tag{20}$$

Hence, a ban on child labor hurts workers if

$$f'\left(\frac{N}{n}\right) < (m\gamma + 1)f'\left(\frac{N + \gamma mN}{n}\right). \tag{21}$$

Clearly we can find parameters under which this inequality may or may not hold. Hence, a ban could hurt worker households and also benefit them.[14]

Let us consider the case where a total ban cannot be implemented. This could be because of difficulties in monitoring. Children can be stopped from laboring in factories but there is little that government can do to stop children laboring on their own family farms. Similarly, Senator Harkin's bill in the United States can conceivably drive child labor out of the export industries in the Third World but can do precious little to prevent child labor in industries which produce for the domestic market. In anticipation of this bill becoming law, the Bangladesh Garment Manufacturers and Exporters Association took steps to fire children from their factories. "The children went from jobs in garment factories to much worse jobs, such as breaking bricks in the hot sun or, even worse, prostitution" (Sarah L. Bachman, 1995 p. 3). Another problem with some of these well-meaning suggestions for intervention is that they can provide a refuge for people and lobbies with other agendas that are not as well meaning, such as protectionism.

---

[13] It is interesting to note that some of the same effects of a ban on child labor can be achieved through the implementation of a minimum-wage law.

[14] We have in our analysis ignored the fact that a small but nonnegligible number of children belong to no family. They are "abandoned," and make their own decision to work or not work. A blanket ban on child labor, without any provision for such children, will almost always work against the interest of these children. A model that explicitly deals with the problem of "street children" (for an empirical account, see William Myers, 1988) would have to be based on very different assumptions from the ones we have used here.

It is therefore important to investigate the effect of partial bans. To model such an intervention, let us introduce an innocuous difference between the n firms in the above model. Suppose n_1 firms are run by redheaded entrepreneurs and $n_2\,(=n-n_1)$ firms by greenheaded ones. Government, we assume, can only administer a ban on the "red" firms. What will be the effect of such a ban?

So we start from a bad equilibrium (w_A^*, w_C^*) and then have a ban announced for the n_1 red firms. The consequence of this depends on the size of n_1.

Note that each firm's demand for effective labor in the equilibrium is $(1+m\gamma)N/n$. Hence, total demand for labor from the red firms is $n_1(1+m\gamma)N/n$. Suppose that the number of red firms, n_1, is so few that the following is true:

$$\frac{n_1(1+m\gamma)N}{n} \leq N. \tag{22}$$

In other words, define $n' \equiv n/(1+m\gamma)$; and suppose $n_1 < n'$. Then the ban has no effect. All the red firms employ adults and the green firms employ the remaining adults and all the mN children.

Now suppose $n_1 > n'$. Evidently the preban demand for labor by the banned red firms exceeds the supply for adult labor. Hence, the pre-ban equilibrium cannot be sustained since we now have an excess demand for adult labor (and excess supply of child labor). Several possibilities arise in this case. One interesting situation would arise if there exists $(\tilde{w}_A, \tilde{w}_C)$ such that

$$\tilde{w}_A \leq s + sm\beta - m\tilde{w}_C; \tag{23}$$

$$f'\left(\frac{N}{n_1}\right) = \tilde{w}_A; \tag{24}$$

$$f'\left(\frac{\gamma mN}{n_2}\right) = \frac{\tilde{w}_C}{\gamma}. \tag{25}$$

Combining (23)–(25), we can equivalently write the following condition:

$$f'\left(\frac{N}{n_1}\right) \leq s + sm\beta - \gamma m f'\left(\frac{\gamma mN}{n-n_1}\right) \tag{26}$$

If n_1 satisfies (26) then we have, after ban, an equilibrium where adult wage is $\tilde{w}_A$ child wage $\tilde{w}_C$. The red firms employ adults, the green firms only children. All dren still work. And since $\tilde{w}_C$ is clearly less w_C^*, child wage is less after the ban. From point of view of banishing child labor, the would, in this case, have to be consider failure. It does not diminish child labor, child wage. This is the possible predicam that one has to worry about in recommen a legislation which can only effect a partial

It is worth noting, however, that even if (is satisfied and the ban is a failure from point of view of controlling child labor, it or may not lower the utility of the wor households. That depends on the following (26) is satisfied and $\tilde{w}_A$ and $\tilde{w}_C$ are such th

$$\tilde{w}_A + \tilde{w}_C < w_A^* + mw_C^*, \tag{27}$$

then the ban not only worsens the child-la condition but it lowers household utility well. If the inequality in (27) is reversed, household utility rises. If, on the other h n_1 is very large, and close to n, it is easy see that the ban works *as if* it were a total and the labor market would settle at the g equilibrium.

The above discussion is at best a surrog analysis of what would happen in a develop country if its exports which use child labor banned. A fuller model can potentially be to address a variety of policy questions in regard. Suppose, for instance, the export in try is competitive and therefore runs on a s der profit margin and this is a small country so it faces a fairly elastic demand. Then a on child labor can increase the cost of pro tion and cause the export industry to sh sharply, leaving the worker households w off. But, for a formal analysis we need to on our simple model more complicated s tures so that such questions can be form taken up.

There are many other kinds of policy—taxes, subsidies, and other restrictions—the effects of which can be checked out using our model.

To sum up, bans are a powerful instrument but by no means unequivocally desirable. One has to be very careful about the empirical context before using this instrument. If there are multiple equilibria in the labor market, a ban is a benign policy intervention and worthwhile. But if the market has only one equilibrium which is likely in very-poor countries, then a ban can worsen the condition of the labor households. Partial bans are especially likely to backfire and cause deterioration in labor conditions. The first-best policy is to attack the problem at its source. This entails improving the condition and scope for adult labor.

V. Concluding Remarks: Fertility and Gender

The model built in this paper has implications for analyzing fertility and population policy. It seems likely that the multiple equilibria in the labor market could bring about a multiplicity of equilibria pertaining to fertility choices of the household, once such choices are endogenized.[15] If our conjecture is right, then this will have implications for the kinds of population policy that we espouse.

Suppose an economy is at an equilibrium where fertility is high and children work. It is pointless in such a situation to send extension workers to households to explain to them the irrationality of large households. This is because the large family is a conscious, rational decision. This is, of course, a widely held view. What is interesting is that even though there is no individual irrationality at this equilibrium, the equilibrium may well be exhibiting *group* irrationality. Everybody would not only be better off if everybody had small families, but every individual family may prefer to be small if other families were small. Hence, the policies we would have to conceive of would attempt to deflect the economy from the high-fertility equilibrium to the low-fertility one.

More generally, the framework developed in this paper, including the diagrammatic technique, should be applicable to several areas other than child labor. Whenever we have two or more variables being supplied by one decision maker, some of the same issues discussed here are likely to crop up. Several gender-related matters belong to this category. Traditional households, where the husband decides not only about his own work but also that of his wife's, may give rise to female labor-supply functions such that we would have multiple equilibria with women being excluded totally from the labor market in some of the equilibria. There is however a caveat to this noted in the next paragraph. If the policy maker does not share the husband's judgment, then she could use this kind of model to decide how best to intervene in the market in order to enhance or curb women's employment. There may also be important issues of gender within the domain of child labor. There is evidence that the labor-supply response of girls and boys to changes in labor-market conditions can be very different (see, e.g., Goldin, 1979). Our model can, in principle, be extended to study the markets of boy-labor and girl-labor.

If we do use this kind of a model to analyze gender issues and, in particular, the supply of female labor, one important real-life difference needs to be kept in mind. There is some evidence that when women begin to work outside the household and contribute to the household's income they also have more influence on household choices and decisions (see, e.g., John Strauss and Duncan Thomas, 1995; Nancy E. Riley, 1997). Hence, it is natural to expect that a man will take this into account when he decides to send his wife out to work. In addition, even if the initial decision about whether a woman works or not belongs to her husband, once she begins working the decision whether she continues to work or not may cease to be the husband's decision. This anticipated shift in decision-making is likely to introduce some important complications to modeling female labor, as distinct from child labor.

---

[15] The link between the market for child labor and fertility behavior has been studied in the literature (see, for instance, Mead Cain and A. B. M. Khorshed Alam Mozumder, 1981) but the possibility of multiple equilibria in this context seems to have been overlooked.

Finally, one important area of practical concern to which models such as these can be brought to bear is the debate on international labor standards. Because of the importance of this topic in international politics, there is now a growing literature commenting on it (see, e.g., Gary S. Fields, 1994; Dani Rodrik, 1995). These are matters which, despite the growing interest, are still discussed without an accepted formal analytical framework. Combining the model of this paper with trade could take us towards a formal framework.

REFERENCES

Anderson, Michael. *Family structure in nineteenth century Lancashire*. Cambridge: Cambridge University Press, 1971.

Ashagrie, Kebebew. "Statistics on Child Labour." *Bulletin of labour statistics*. Geneva: International Labour Organization, 1993-3, pp. 11–24.

Bachman, Sarah L. "Children at Work" (Commentary). *San Jose Mercury News*, July 16, 1995, p. 3.

Bequele, Assefa and Boyden, Jo, eds. *Combating child labour*. Geneva: International Labour Organization, 1988.

Browning, Martin; Bourguignon, François; Chiappori, Pierre-André and Lechene, Valérie. "Income and Outcomes: A Structural Model of Intrahousehold Allocation." *Journal of Political Economy*, December 1994, *102*(6), pp. 1067–96.

Cain, Mead and Mozumder, A. B. M. Khorshed Alam. "Labour Market Structure and Reproductive Behaviour in Rural South Asia," in Gerry Rodgers and Guy Standing, eds., *Child work, poverty, and underdevelopment*. Geneva: International Labour Organization, 1981, pp. 245–87.

Cunningham, Hugh. "The Employment and Unemployment of Children in England c. 1680–1851." *Past and Present*, February 1990, *126*, pp. 115–50.

Fields, Gary S. "Labor Standards and International Trade." Mimeo, Cornell University, 1994.

Goldin, Claudia. "Household and Market Production of Families in a Late Nineteenth Century American Town." *Explorations in Economic History*, April 1979, *16*(2), 111–31.

Grootaert, Christiaan and Kanbur, Ravi. "C Labour: An Economic Perspective." *Inte national Labour Review*, 1995, *134*(2), 187–203.

Gupta, Manash R. "Wage Determination o Child Worker: A Theoretical Analysi *Review of Development Economics*, 19 (forthcoming).

Horrell, Sara and Humphries, Jane. " 'The E ploitation of Little Children': Child La and the Family Economy in the Industr Revolution." *Explorations in Econo History*, October 1995, *32*(4), pp. 48 516.

Moehling, Carolyn M. "The Intrahouseho Allocation of Resources and the Particip tion of Children in Household Decisio Making: Evidence from Early Twenti Century America." Mimeo, Northweste University, 1995.

Myers, William. "Alternative Services for Str Children: The Brazilian Approach," i Assefa Bequele and Jo Boyden, eds., *Co bating child labour*. Geneva: Internatio Labour Organization, 1988, pp. 125–43.

Nardinelli, Clark. *Child labor and the Ind trial Revolution*. Bloomington, IN: India University Press, 1990.

Riley, Nancy E. "Gender, Power, and Popu tion Change." *Population Bulletin*, M 1997, *52*(1), pp. 2–46.

Rodrik, Dani. "Labor Standards and Inte national Trade: Moving Beyond the Rh oric." Mimeo, Columbia Universit 1995.

Sen, Amartya K. "Gender and Cooperati Conflict," in Irene Tinker, ed., *Persist inequalities: Women and world develo ment*. New York: Oxford University Pr 1990, pp. 123–49.

Silvers, Jonathan. "Child Labor in Pakistan *Atlantic Monthly*, February 1996, pp. 7 92.

Strauss, John and Thomas, Duncan. "Hum Resources: Empirical Modeling of Hou hold and Family Decisions," in Je Behrman and T. N. Srinivasan, eds., *Ha book of development economics, volu IIIA*. Amsterdam: Elsevier, 1995, pp. 188 2023.

Udry, Christopher. "Gender, Agricultural Production and the Theory of the Household." *Journal of Political Economy*, October 1996, *104*(5), pp. 1010–46.

Vincent, David. *Bread, knowledge and freedom: A study of nineteenth-century working class autobiography*. London: Europa Publications, 1981.

Weiner, Myron. *The child and the state in India: Child labor and education policy in comparative perspective*. Princeton, NJ: Princeton University Press, 1991.

A Theory of Holdouts in Wage Bargaining

By Wulong Gu and Peter Kuhn*

Holdouts (the continuation of negotiations beyond the contract expiry date) are the most common form of disputes in labor contract negotiations. We model holdouts as a delaying tactic employed by unions to obtain information about other bargaining outcomes in their industry. Novel implications of our model include a positive association between holdout duration and the number of bargaining pairs negotiating contracts simultaneously; bunching of holdout durations within these "negotiating groups"; and fewer strikes among holdouts which end later in groups. Using a large panel of contract negotiations in Canadian manufacturing, we find considerable support for these predictions. (*JEL* J52)

Understanding the causes of disputes in union bargaining has long been a major focus of research in labor economics, both theoretical (e.g., Beth Hayes, 1984) and empirical (e.g., David Card, 1988). Until recently, however, a common limitation of most theoretical models used in this area has been the tendency to ignore a potentially important option of the bargaining parties. That option is simply to fail to agree on a new contract, but to continue working under the terms of an expired one (i.e., to "hold out"), rather than proceeding immediately to a strike or lockout. Since holdouts are the most common form of disputes in U.S. union contract negotiations (Peter C. Cramton and Joseph S. Tracy, 1992), and since the option to hold out can dramatically change the other outcomes of bargaining models (such as strikes and wages),[1] an understanding of holdouts would seem to be an essential component of any understanding of union contract negotiations.

To date, the only existing formal model of holdouts treats them as an alternative threat a union can impose on a firm, sometimes more appropriate to attaining the union's ends than a full-fledged strike (Cramton and Tracy, 1992, 1994). In this paper we develop and test an alternative model of holdouts. A key distinguishing feature of our model is its focus on the interactions between multiple bargaining pairs in an industry; in contrast to Cramton and Tracy, we model holdouts as a delaying tactic employed by unions to obtain information from other bargaining outcomes in their industry. More specifically, if unobservable aspects of profitability are correlated across firms, a union can acquire valuable information about its own firm's ability to pay by observing the bargaining outcomes of other unions (Kuhn and Gu, 1996). When a number of contracts expire at the same time, an information externality, similar to those in recent sequential learning models (e.g., Rafael Rob, 1991; Abhijit V. Banerjee, 1992) arises; each union has an incentive to delay its settlement and let the other unions conclude their negotiations first.[2]

Like Cramton and Tracy's, our model of holdouts predicts that holdouts should be

---

* Gu: Micro-Economic Analysis Directorate, Industry Canada, Room 527G, 5th Floor West, 235 Queen St., Ottawa K1A 0H5, Canada; Kuhn: Department of Economics, McMaster University, 1280 Main St. West, Hamilton, Ontario L8S 4M4, Canada. Research on this paper was supported by the Social Sciences and Humanities Research Council of Canada. We thank Ken Binmore, Tilman Börgers, Martin Osborne, Arthur Sweetman, participants at seminars and conferences, and two anonymous referees for helpful comments. Special thanks go to Alan Harrison for providing the data.

[1] See, for example, Raquel Fernandez and Jacob Glazer (1991).

[2] In a recent paper, Christophe Chamley and Douglas Gale (1994) investigate strategic delay and informational externalities in a model of investment. Daniel L. McFadden and Kenneth E. Train (1996) do so in a model of consumption, and Gu (1996) in a model of the adoption of innovations. The pure informational externality in these models is similar to the one in our model.

shorter and less frequent when the wage settlement in the existing contract is lower, and when unions are more optimistic about the firm's ability to pay. More novel features of our model include the following. First, it is capable of generating holdouts which are followed by strikes (in our data, strikes almost never occur in the absence of a prior holdout). Second, while allowing for the possibility, the model does not require the ad hoc assumption of a production inefficiency during holdouts to generate holdouts. Third, the model has a number of novel predictions, including the following. Since the information externality identified above is exacerbated by the addition of more bargaining pairs, mean holdout durations should increase with the number of bargaining pairs negotiating simultaneously in an industry. Since the incentive to outwait one's peers declines as more of their holdouts have already ended, holdout expiry dates should exhibit a kind of "bunching" within such "negotiating groups." And since bargaining pairs who end their holdouts later should have access to better information, holdouts which end later in a group should be less likely to end in strikes.

We examine the above implications using a large sample of contract negotiations in Canadian manufacturing industries for the period from 1965 to 1988. While not necessarily inconsistent with other models of holdouts, the results are remarkably consistent with our view of holdouts as a "waiting game" that is played among a group of unions. The context of other negotiations in one's industry—a context which has been relatively neglected in the economic analysis of bargaining and strikes to date—thus appears to be an important determinant of union bargaining outcomes.[3]

Sections I and II of this paper analyze a model with two bargaining pairs, in which the bargaining pairs negotiate their contracts simultaneously. The model illustrates unions' incentives to delay their wage settlements (or hold out), in order to gain better information about their firms' ability to pay by observing the bargaining outcomes in other unions. To determine how long to hold out, unions trade off the potential benefits of learning from others against the cost associated with delayed wage settlements. Section III extends the model to the case of more than two bargaining pairs, and sets out the main testable implications that distinguish the current model from other views of holdouts. Sections IV and V go on to assess the consistency of these implications with the Canadian contract and strike data. Section VI concludes.

I. A Model

Consider two bargaining pairs: bargaining pair one (union 1 and firm 1) and bargaining pair two (union 2 and firm 2). The existing contracts for both pairs have expired at the same date, and both are negotiating the wage to be paid during a contract of duration T. Informally, the game between the two unions and firms can be described as follows: at each instant, each union decides whether to start to play a negotiation game with its firm, conditional on its not having started by that date. To focus on the main issue of the paper—the incentive by unions to hold out in order to learn about their firms' ability to pay from bargaining outcomes in other unions—we shall use a very simple negotiation game. In this game, the firm has private information about its ability to pay, and the union makes a take-it-or-leave-it offer, to which the firm responds. This negotiation game has been analyzed by Drew Fudenberg and Jean Tirole (1991), among others.[4]

---

[3] For example, almost all formal economic models of strikes (e.g., Hayes, 1984; Card, 1990) consider only a single bargaining pair in isolation. While empirical analyses of strikes have occasionally tried to control for past bargaining outcomes of the *same* bargaining pair (e.g., Card, 1988), to our knowledge no such studies have considered the effects of concurrent or past behavior of other bargaining pairs within the industry.

[4] In other work (Kuhn and Gu, 1996) we have modelled the effects of information spillovers across bargaining pairs in an infinite-horizon model, where the union can make any number of offers to the firm. Unlike the current example, which is simplified to allow us to endogenize the sequencing of negotiations, such a model generates predictions about strike duration as well as incidence. To date, we have not considered the effects of alternating-offer models of bargaining (e.g., Fernandez and Glazer, 1991), which are very hard to extend to the case of multiple bargaining pairs with correlated unobserved shocks. Informally, however, since there are potential efficiency gains

TABLE 1—JOINT PROBABILITIES OF "GOOD" AND "BAD" STATES AT FIRMS 1 AND 2

Firm 1's state		Firm 2's state		
		Good	Bad	
	Good	p_{11}	p_{12}	g_1
	Bad	p_{21}	p_{22}	b_1
		g_2	b_2	

Formally, the game is described as follows. Time is continuous from 0 to T, and the discount rates for both unions and firms are r. First, at (or before) time 0 (the expiration date of the previous contracts), nature determines whether the firms are in the good or bad state. Their profits gross of labor costs are $\pi_g > \pi_b > 0$ in these two states respectively. The firms know their own states, but the unions do not. However, the unions share a common prior distribution over the joint probabilities of the states in the two firms, which is described in Table 1: the probability that firm 1 is in the good state, g_1, equals the sum of the joint probabilities p_{11} and p_{12}; b_1, b_2 and g_2 are defined analogously. The conditional probabilities corresponding to Table 1 are:

$$(1) \quad P(2G|1G) = p_{11}/(p_{11} + p_{12});$$

$$(2) \quad P(2G|1B) = p_{21}/(p_{21} + p_{22});$$

$$(3) \quad P(1G|2G) = p_{11}/(p_{11} + p_{21});$$

$$(4) \quad P(1G|2B) = p_{12}/(p_{12} + p_{22}),$$

where $1G$ indicates a good state in firm 1, etc. We assume, without loss of generality, that firm 1 is at least as likely to be in the good state as firm 2, i.e., $g_1 \geq g_2$; and we restrict our attention to the case of positive correlation between the firms' profits. By definition, the two firms' profits are positively correlated if and only if

$$(5) \quad P(jG|iG) > g_j, \text{ and } P(jB|iB) > b_j \text{ for } i \neq j = 1, 2.$$

This simply means that after observing the other firm i to be in the good (bad) state, union j becomes more optimistic (pessimistic) about its firm's profit and hence updates its prior upwards (downwards).

After the states of the two firms are determined, at each instant each union decides whether to make a wage demand, conditional on its having not made a demand before that instant. If a union decides to make a wage demand at time t $(0 \leq t \leq T)$, it then chooses the magnitude of the wage demand $d \in [0, \infty)$, to which the firm responds immediately by accepting or rejecting that wage demand. Acceptance means exchange takes place at the wage demanded by the union, with the union receiving the wage demand d and the firm receiving its profit net of the wage demand for the remaining period $[t, T]$; rejection means both parties get their alternative utility level for the remaining period $[t, T]$; this level is normalized to zero. During the course of contract negotiations, each union observes what is going on in the contract negotiations of the other bargaining pair. Specifically, it observes whether the other union has already made a demand, whether the firm has accepted, and the magnitude of the wage settlement if the firm has accepted. In choosing its strategy, we assume in this paper that each union acts purely in its own self-interest. We therefore abstract from the possibility of meaningful collusion between unions or, for example, between locals of the same national union in our formal model.[5]

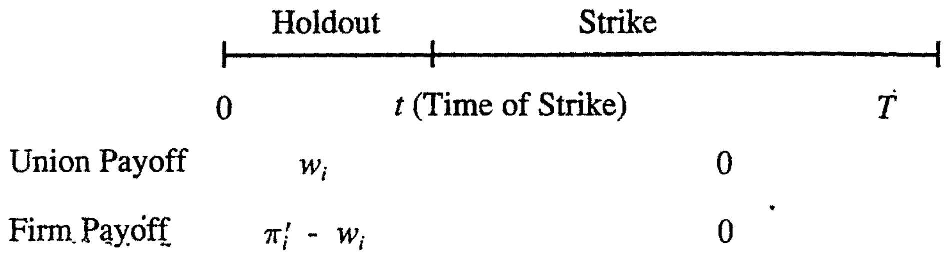
To conclude the specification of the game, we need to specify the payoffs of both parties for the period from the expiration of the previous contracts to the time when unions make a wage demand. That period is defined as a

---

to be shared by workers and firms by outwaiting other bargaining pairs, one would expect the same factors (e.g., number of bargaining pairs) to affect average holdout length even in an expanded model which allowed for multiple offers, and/or alternating offers.

[5] Kuhn and Gu (1998) explore the effects of union (and firm) collusion on strikes in a sequential bargaining model with information spillovers. Empirically, we interpret this paper's evidence that holdouts increase with negotiating group size, and that this relationship is similar whether we look at bargaining units in the same or in different national unions (see footnote 24) as evidence of imperfect collusion among union locals.

(a) Payoff if a Strike Occurs



(b) Payoff if No Strike Occurs

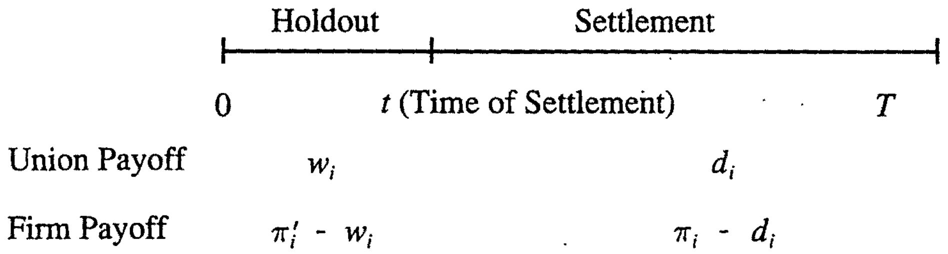


FIGURE 1. UNION AND FIRM PAYOFFS IN BARGAINING PAIR i

Note: d_i: wage demand by union i. π: firm i's profit after a settlement. π_i': firm i's profit during a holdout.

holdout in Cramton and Tracy (1992). In the United States, according to the National Labor Relations Act, the terms and conditions of employment during holdouts are governed by the previous contract. In Canada, similar laws apply. Therefore, during a holdout, a union is paid the wage rate in the previous contract and a firm receives its profit net of the wage cost; specifically, union 1 is paid w_1 and union 2 is paid w_2.

The payoff flows of both parties in bargaining pair i ($i = 1, 2$) are summarized in Figure 1. Note that during the period of a holdout, a firm's profit π_i' might be lower than its profit π_i in the normal production period after the conclusion of contract negotiations. Cramton and Tracy (1992) give a number of potential reasons for such inefficiencies: for example, workers may be more likely to "work to rule" during a holdout; potential customers and suppliers may be reluctant to deal with firms if a holdout signals a greater likelihood of an impending strike; and a holdout may delay changes in work rules which might increase productivity. It is worth noting, however, that unlike earlier models, these inefficiencies are not required to generate holdouts in the current model: positive holdouts will occur here even when $\pi_i = \pi_i'$.

For the remainder of the paper we shall assume that wage settlements in the previous contracts are lower than the firms' profit in the bad state π_b, i.e., $w_i < \pi_b$ for $i = 1, 2$. Otherwise, if the existing wage settlements are greater than the firms' profits in the bad state ($w_i > \pi_b$ for $i = 1, 2$), unions have no incentive to make a wage demand and conclude contract negotiations when they are sufficiently pessimistic about the firms' ability to pay.[6]

---

[6] Situations in which w_i is consistently less than π_b are of course more likely when there is considerable real wage growth, and/or considerable inflation in the economy, and when nominal wage contracts are of relatively long duration. Broadly speaking, this was indeed the case for most of our sample period. In an expanded model, both firms *and* unions will prefer the likely outcome of a new contract to the old one when contract wages have efficiency effects, and when evolving economic conditions make old contract provisions inefficient.

An equilibrium of the above game consists of unions' strategies, firms' strategies, and unions' beliefs. A union's strategy consists of the timing and magnitude of its wage demand—specifically, at each instant a union decides whether to make a wage demand and the magnitude of its wage demand, conditional on its having not done so before that instant. A firm's strategy is simply its acceptance decision. In a perfect Bayesian equilibrium, we must have: (1) the strategies of all players are optimal; and (2) the unions' beliefs are consistent with Bayes' law for any possible history of the game whenever it is applicable.

We now proceed to characterize the perfect Bayesian equilibrium of the game. We shall do so via two steps. First, taking as given the *timing* of both unions' wage demands, we derive the unions' optimal wage demands and the firms' responses (Lemma 1); we then note that the game between the two unions in choosing the timing of their wage demands is a war-of-attrition game. We use this fact to determine the timing of the unions' wage demands (Propositions 1, 2, and 3).

LEMMA 1:

1. *Each firm accepts a wage demand if and only if it is less than or equal to its profit;*
2. *If the two unions decide to make wage demands at the same time t, union i ($i = 1, 2$) demands π_g if $g_i > \pi_b/\pi_g$ ($\equiv b$), and demands π_b if $g_i < b$;*[7]
3. *If union i makes a wage demand before union j, then union i demands π_g if $g_i > b$ and π_b if $g_i < b$; union j demands π_g if $p_j > b$ and π_b if $p_j < b$, where p_j is the posterior belief of union j after observing the outcomes in union i. It is calculated as follows: $p_j = g_j$ if the wage settlement in union i is π_b; $p_j = P(jG|iG)$ if the high wage settlement π_g is reached without a strike in union i; $p_j = P(jG|iB)$ if a strike occurs in union i.*

PROOF:
See Appendix A.

When unions make wage demands at different times, unions who make demands later have an opportunity to learn from the outcomes in previous negotiations. As demonstrated in Kuhn and Gu (1996), the opportunity to learn from others raises unions' utilities and reduces strikes. Intuitively, this follows from the fact that unions' learning improves their information about firms' ability to pay. Therefore, when two contract negotiations proceed simultaneously in an industry, each union has an incentive to hold out and let the other union conclude negotiations first, thus learning from the outcomes in the other union. However, by so doing, the union is paid the existing lower wage during the period of holdout and has to delay a possible higher wage settlement. Unions trade off the benefit of the opportunity to learn from others against the cost associated with delayed higher wage settlements to decide the timing of their wage demands.

Propositions 1, 2, and 3 characterize the optimal timing and magnitude of the unions' wage demands in the perfect Bayesian equilibrium for various values of the two unions' priors g_1 and g_2.

PROPOSITION 1: *If $g_1 < b$ and $g_2 < b$, both unions demand π_b immediately at time* 0.

PROOF:
See Appendix A.

In this case where both unions are sufficiently pessimistic about firms' ability to pay, they make low wage demands of π_b immediately after the expiration of the previous contracts. Both firms will always accept. Neither holdouts nor strikes occur.

PROPOSITION 2: *If $g_1 > b$ and $g_2 < b$, in the unique perfect Bayesian equilibrium the two unions' optimal strategies must have the following features:*

1. *Union 1 makes a high wage demand of*
   at time 0;
2. *When $P(2G|1G) < b$, union 2 makes*

---

[7] If $g_i = b$, union i is indifferent between π_g and π_b. Throughout the paper we will ignore this borderline case.

low wage demand of π_b at time 0; when $P(2G|1G) > b$, it makes a wage demand immediately after union 1, with the magnitude of the demand depending on the outcome in union 1: it demands π_g if union 1 does not strike and demands π_b if union 1 strikes.

PROOF:

See Appendix A.

The most interesting case is in Proposition 3, when $g_1 > b$ and $g_2 > b$. In that case if the two firms' abilities to pay are highly correlated such that $P(1B|2B) > 1 - b$ and $P(2B|1B) > 1 - b$, there exist many perfect Bayesian equilibria. However, the one equilibrium involving randomization by both unions seems to be the natural candidate in the present context where unions have strategic uncertainty about the timing of the other unions' wage demands (see also Chamley and Gale, 1994). Furthermore, as Martin J. Osborne (1985) has shown, that is the only equilibrium that satisfies a particular version of trembling-hand perfection.[8] In what follows, we shall focus on this particular equilibrium only. In this equilibrium, both unions adopt mixed strategies in choosing the time of their wage demands, and union i makes a wage demand between time t and $t + dt$ with probability $h_i(t)dt$, conditional on neither union having made an offer at time t. Thus $h_1(t)$ and $h_2(t)$ are hazard rates for making the first offer in the industry.

PROPOSITION 3: *When $g_1 > b$ and $g_2 > b$, the equilibrium strategies of the two unions must have the following features:*

1. *If the firms' profits are highly correlated such that $P(2B|1B) > 1 - b$ and $P(1B|2B) > 1 - b$, union i makes a wage demand with probability $h_i(t)dt$ between time t and $t + dt$, conditional on neither having made a demand at t, where:*

$$(6)\ h_i(t) = \frac{r(g_j\pi_g - w_j)}{(1-g_i)\pi_g[(b - P(2G|1B)]} \times \frac{1}{1-e^{-r(T-t)}},$$

and $i, j \in \{1, 2\}, i \neq j$.

The union which makes a demand first demands π_g; immediately after that union has made a wage demand, the other union demands π_b if a strike occurs in the preceding negotiation, and demands π_g if no strike occurs in the preceding negotiation.[9]

2. *If the firms' profits are highly uncorrelated such that $P(1B|2B) < 1 - b$ and $P(2B|1B) < 1 - b$, both unions demand π_g immediately at time 0.*
3. *If the degree of correlation between the firms' profits is in an intermediate range such that $P(1B|2B) < 1 - b$ and $P(2B|1B) > 1 - b$, union 1 demands π_g at time 0 and union 2 makes a wage demand immediately afterwards. It demands π_g if union 1 doesn't strike and demands π_b if union 1 strikes.*

Note by the assumption $g_1 > g_2$, we have $P(2B|1B) > P(1B|2B)$.

PROOF:

See Appendix A.

One simple prediction of our model follows directly from (6). Note that whenever unions adopt mixed strategies (i.e., whenever there are positive holdouts in the current model), the hazard rates $h_1(t)$ and $h_2(t)$ are increasing functions of time: as time proceeds, the (discounted) benefits of learning from others to obtain high wage settlements decrease and thus unions are more likely to make wage demands. It is also interesting to note that, in all cases where the equilibrium does *not* involve randomization, the unions either move simultaneously or union 1 (which has the higher *ex ante* prior of its firm being in the good state) moves first. This accords with the notion in the industrial relations literature that unions "target" high-profitability firms in an industry first.

---

[8] In Osborne (1985), an equilibrium is trembling-hand perfect if the equilibrium strategies are robust with respect to a perturbation of these strategies that incorporates the probability of small mistakes that players concede at time 0.

[9] As mentioned, in this case, there exist many perfect Bayesian equilibria. One such equilibrium is as follows: union 1 makes a wage demand at time 0; at each instant, union 2 makes a wage demand if and only if the other union has made a demand. By the same argument as in Osborne (1985), we can show that none of these equilibria are trembling-hand perfect because the equilibrium strategies are not robust with respect to a perturbation of these strategies that incorporates the probability of small mistakes that unions may make wage demands at time 0.

If the firms' profits are perfectly correlated, i.e., $p_{12} = p_{21} = 0$ in Figure 1, we have the following corollary (denote p_{11} by p: the probability that both firms are in the good state).

COROLLARY 1: *If the firms' profits are perfectly correlated, the unions' strategies must have the following features:*

1. *If $p < b$, both unions make a low wage demand of π_b immediately at time 0;*
2. *If $p > b$, conditional on neither union having made a wage demand at t, union i makes a wage demand between t and $t + dt$ with probability $h_i(t)dt$, where*

$$(7)\quad h_i(t) = \frac{r(p\pi_g - w_j)}{(1-p)\pi_b}\cdot\frac{1}{1-e^{-r(T-t)}};$$

$$i, j \in \{1, 2\},\ i \neq j.$$

The union which moves first demands π_g. After a union makes a demand, the other union makes a demand immediately afterwards: it makes a high wage demand of π_g if there was no strike in the preceding negotiation and makes a low wage demand of π_b if there was a strike.

II. Comparative Statics

In this section we investigate how holdout durations vary with the parameters of the model, in particular the wage settlements in the previous contracts and the probability that firms are in the good state. For simplicity, we restrict our attention to the case where the firms' profits are perfectly correlated, and to situations where unions are sufficiently optimistic about the firms' profits for holdouts to exist in equilibrium ($p > b$).

PROPOSITION 4: *When $p > b$:*

1. *The higher are the existing wage settlements, the longer are holdout durations;*
2. *The lower is the probability that firms in the good state, the longer are holdo durations;*
3. *The longer is the contract duration, t longer are holdout durations.*

PROOF:

See Appendix A.

When the wage settlements in the previo contracts increase, the cost of delaying wag settlements decreases, which leads to longe holdouts. Similarly, when unions believe it less likely that firms are in the good state, the expect lower wage settlements, and the cost o delaying is then lower. This causes longer holdout durations. Finally, when contract duratio increase, the discounted benefit of learnin from others to obtain higher wage settlemen increases. Unions then have a greater incentiv to delay wage settlements.

III. N ($N > 2$) Bargaining Pairs

In this section we extend the model in Section II to the case of n ($n > 2$) bargaining pai and see how expected holdout durations va with the number of bargaining pairs in an industry. For simplicity, we assume that th firms' profits in the n bargaining pairs are perfectly correlated. With probability p all n firm are in the good state; and with probability $1 - p$ all n firms are in the bad state. As befor the firms' profits gross of labor cost are π_g an π_b in those two states respectively. Finally, w assume that wages in all previous contracts a w. Proposition 5 identifies the unique symmetric equilibrium of the game.[10]

PROPOSITION 5: *In the unique symmetric perfect Bayesian equilibrium, the unions' strategies must have the followin features:*

1. *If $p < b$, all n unions make the low wa demand of π_b immediately at time 0;*

---

[10] As in the case of Proposition 3, there also exist asymmetric equilibria, but these do not satisfy a version trembling-hand perfection which is natural for this co (Osborne, 1985).

(2) If $p > b$, conditional on none of the unions having made a demand at t, each union makes a wage demand between time t and $t + dt$ with probability $h(t)dt$, where

$$h(t) = \frac{r(p\pi_g - w)}{(n-1)(1-p)\pi_b} \cdot \frac{1}{1 - e^{-r(T-t)}} \tag{8}$$

$$\equiv \frac{1}{(n-1)} A(t).$$

The union which makes a wage demand first demands π_g. Immediately after a union has made a wage demand, all other unions make wage demands: they make the high wage demand π_g if there was no strike in the preceding negotiation and make the low wage demand π_b if there was a strike.

PROOF:

See Appendix A.

In equation (8), a higher number of negotiations, n, leads to a lower hazard $h(t)$ at every instant $t \in [0, T]$. As the number of negotiations in an industry increases, each union thus has a greater incentive to delay making the first wage demand. Intuitively, as the number of negotiations in the model expands, a union's expected benefit of delaying its wage demand rises because it expects a higher probability that at least one of the other unions will make a demand during that period of delay, thus learning about its firm's ability to pay.

Given n unions all making the first wage demand with probability $h(t)dt$, and given the result that all other unions make a demand immediately after the first one, the overall hazard rate for holdouts ending in each of the n firms is just given by:

$$H(t) = \frac{n}{(n-1)} A(t), \tag{9}$$

which is also decreasing with n, but at a diminishing rate. Intuitively, as the number of unions expands, the probability of each union making the first offer falls at a rate proportional to $1/(n - 1)$, which depends on the number of *other* unions $(n - 1)$ from whom useful information might be gathered by waiting. At the same time, the total number of unions making offers with probability $h(t)dt$ expands at a rate proportional to n. Since the former effect dominates the latter [$n/(n - 1)$ falls with n], we expect average holdout duration (which is inversely related to the hazard) to increase with n, but at a decreasing rate.[11]

In sum, the model of holdouts proposed in this paper has at least three novel implications stemming from the information externalities across bargaining pairs in an industry. First, and most directly relevant to the question of which factors explain the existence and severity of holdouts, average holdout duration should increase with the number of bargaining pairs negotiating contracts at the same time: the more "peers" one can potentially learn from by waiting, the greater the incentive to delay one's own negotiations.[12] Second, note that, at least in the simple models examined in this paper (with either only two bargaining pairs or perfect correlation of the unobserved shock among firms), all the other bargaining pairs in the industry end their holdouts immediately after the first pair has ended theirs. Third, note that (because all the relevant information is publicly available at that time), no strikes ever occur among these "following" bargaining pairs. While we do not expect these predictions to be literally true in a more complex world, we should still expect to see both a "bunching" of holdout expirations within negotiating groups and a lower rate of strike incidence among holdouts which end later in a group if information spillovers are an important explanatory factor in union bargaining. We test these implications using a large sample of contract negotiations in

---

[11] Another way to understand this result, suggested by a referee, is to note, from (8), that indifference between making a wage demand now and waiting implies that the total "probability" of any *other* union making an offer, $(n - 1)h$, must be independent of the number of negotiating units, n.

[12] Although a greater number of bargaining pairs increases holdouts in the current model, it may be worth noting that, because holdouts occur even when $\pi_i = \pi_i'$, this does not necessarily entail an efficiency cost. Indeed, because strikes are predicted to fall with group size, n, raising n will tend to have positive overall efficiency effects unless the cost of holdouts is very high.

Canadian manufacturing industries in the following two sections of the paper.

IV. Data

The data used in this paper come from two sources: Labour Canada's Wages File, and their Work Stoppages File. Labour Canada's Wages File provides information on the settlement, effective, expiry date, and other characteristics of all Canadian collective bargaining contracts involving 500 or more workers from 1965 to 1988.[13] The Work Stoppages File supplies data on start and end dates of strikes for a broader sample of negotiations. Combining information from these two sources, we define holdouts as the period between the expiry date of old contracts and settlement date of new contracts if no strike occurs, or the period between the expiry date of old contracts and the time a strike begins if a strike occurs. Details of the merging process are provided in Alan Harrison (1994).[14]

In this paper, we focus on holdouts in manufacturing industries, within which it is possible to generate particularly long series of negotiations within fairly detailed, consistently defined industry groups. In total, the Wages File contains data on 2,888 wage negotiations, which comprise 82 three-digit industries over this period. After excluding industries with fewer than 10 contract negotiations over the entire period, plus the "miscellaneous manufacturing" category, we are left with 2,747 negotiations in 55 three-digit industries as our basic sample for analysis.[15]

Some basic descriptive statistics on our sample of negotiations are provided in Table 2. In Table 2, and in all the following analysis, we single out one particular industry—pulp and paper mills—for separate treatment, because of its very distinct bargaining organization, involving a much larger number of negotiations occurring simultaneously than any other industry.[16] While the pulp and paper industry provides a large sample of negotiations occurring in very large groups, separate examination of this industry (which constitutes about 14 percent of the negotiations in our sample) ensures that our results are not driven solely by potentially unmeasured differences in the characteristics or the underlying parameters between pulp and paper mills and all other industries.

Together with Appendix Table B1 (see Appendix B), which provides more detail on the characteristics of negotiations by industry, Table 2 shows the following. First, for the period 1965–1988 the average dispute (strike and/or holdout) incidence is 85.0 percent and the mean (median) dispute duration (conditional on a dispute occurring) is 88.7 (66.6) days. These are considerably higher than those reported by Cramton and Tracy (1992) for the United States, which are 57 percent and 65 (37) days, respectively, for the period 1970–1989. Secondly, holdouts are common, and considerably more common than strikes. For the period between 1965 to 1988, the incidence of holdouts only is 64.0 percent, while the incidence of holdouts followed by strikes are 16.8 percent, yielding a total holdout incidence of 80.8 percent, compared to 21.0 percent for strikes. Third, four of every five [1 − (4.2/21.0)] strikes in Canada began after a holdout, rather than immediately upon contract expiration.[17]

Fourth, as shown in Table B1, there is considerable variation in holdout incidence and

---

[13] Some data are actually available for 1964, but coverage is incomplete. Thus we started our analysis in 1965.

[14] Harrison's merging process also removed a small number of "duplicate" negotiations, defined as agreements involving the same firm and same union, with the same wage settlement and expiry date, from the data.

[15] The analysis was also performed for all industries, and for industries with 20 or more negotiations only, with almost identical results.

[16] This feature of pulp and paper mills has been noted by other analysts (e.g., John C. Anderson, 1989 p. 212), though to our knowledge we are the first to quantify the difference between it and other industries. Another distinct feature of this industry is the high number of bargaining pairs involving locals of the same national union, the Canadian Paperworkers Union. According to our estimates from other industries, however, the latter feature does not appear to have a significant effect on holdout durations (see footnote 24).

[17] Seventy-five percent of strikes began after holdouts lasting more than 2 days, and 70 percent after holdouts lasting more than 7 days, compared to 50 and 39 percent, respectively, for the United States (Cramton and Tracy, 1992 p. 103). Legislation requiring mandatory conciliation and a short "cooling off" period before a strike may be partly responsible for the higher numbers in Canada (Morley Gunderson et al., 1989), but the statistics for both countries cast some doubt on Cramton and Tracy's assumption that strikes either occur at contract expiration or not at all.

TABLE 2—DESCRIPTIVE STATISTICS

	All industries (N = 2747)		All industries excluding paper (N = 2375)		Paper only (N = 372)	
(a) Disputes	Incidence (percent)	Mean (median) duration (days)[a]	Incidence (percent)	Mean (median) duration (days)[a]	Incidence (percent)	Mean (median) duration (days)[a]
Holdouts only[b]	64.0	81.6 (61.0)	62.2	72.6 (55.0)	75.3	128.8 (121.0)
Strikes only[c]	4.2	58.1 (44.8)	4.8	58.0 (44.8)	0.3	67.2 (67.2)
All holdouts[d]	80.8	79.6 (60.0)	78.7	70.7 (54.0)	94.3	126.8 (124.0)
All strikes[e]	21.0	52.5 (36.4)	21.3	48.3 (32.2)	19.4	81.7 (67.2)
All disputes[f]	85.0	88.7 (66.6)	83.5	79.0 (60.0)	94.6	143.2 (129.0)

(b) Contract characteristics	Mean	Percentiles 10th	Percentiles 90th	Mean	Percentiles 10th	Percentiles 90th	Mean	Percentiles 10th	Percentiles 90th
Number of concurrent negotiations[g]	3.2	1	5	1.4	1	2	14.6	2	25
Contract duration (months)	27.5	12	36	27.4	12	36	28.4	24	36
Number of employees covered	1339	500	2180	1377	500	2200	1092	515	1800

[a] Conditional on a dispute occurring. All durations are measured in calendar (not working) days.
[b] Includes only holdouts not followed by a strike.
[c] Includes only strikes not preceded by a holdout.
[d] Includes all holdouts.
[e] Includes all strikes.
[f] Includes holdouts and/or strikes.
[g] Number of concurrent negotiations in same three-digit industry.

duration across industries. Across three-digit industries the incidence of holdouts ranges from a high of 100 percent in the paper box and bag industry to a low of 41.7 percent in commercial printing. The average duration of holdouts ranges from 145.9 days in miscellaneous chemicals to 27.6 in automotive fabric accessories. In pulp and paper mills, holdout incidence (at 94.3 percent) is considerably higher, and mean holdout duration (at 129 days) is almost double the average for all other industries, suggesting that its larger number of concurrent negotiations might indeed be generating longer and more frequent holdouts.

Fifth, the average contract negotiation covered a little over 1,300 workers, and the mean duration of negotiated contracts was about 27½ months, with quite a narrow range (20.3 to 34.7 months) across three-digit industries. On these dimensions, as well as in strike incidence, pulp and paper mills (henceforth, paper) are not very different from all other industries. Bargaining units are somewhat, but not dramatically, smaller in paper, while the differences in strike-proneness and contract duration are very minor.

Finally, consider the number of concurrent negotiations for each negotiation in our sample, which we shall refer to as the size of its "negotiating group." In this paper we define negotiating groups as those with the same (calendar) expiry date for the previous contract within the same three-digit industry.[18] Table 2

---

[18] Much of the analysis was replicated defining all previous contracts ending in the same *month* in the industry as concurrent. It was also replicated using two-digit industries rather than three-digit industries as the relevant industrial context (see an earlier version of this paper, available from the authors). In both cases, the results were very similar, though generally weaker in the two-digit industry case.

TABLE 3—HOLDOUTS AND THE NUMBER OF CONCURRENT NEGOTIATIONS:[a] RAW DATA

All industries				
Number of concurrent negotiations	1	2	3–10	11+
Holdout incidence	78.2 (0.9)	82.3 (2.0)	83.3 (2.3)	98.2 (0.9)
Mean holdout duration (days)[b]	68.8 (1.8)	78.7 (4.3)	108.0 (5.7)	126.5 (4.5)
All industries excluding paper				
Number of concurrent negotiations	1	2	3–4	5+
Holdout incidence	78.1 (1.0)	81.8 (2.2)	83.2 (3.5)	73.5 (5.4)
Mean holdout duration (days)[b]	67.1 (1.7)	76.3 (4.3)	92.5 (7.5)	108.7 (13.2)
Paper only				
Number of concurrent negotiations	1–4	5–20	21–22	23+
Holdout incidence	88.5 (3.3)	93.5 (2.5)	98.1 (1.3)	97.3 (1.9)
Mean holdout duration (days)[b]	138.8 (10.4)	107.2 (9.1)	146.1 (7.2)	108.2 (6.3)

Notes: Standard errors in parentheses.

[a] Within three-digit industries.

[b] Conditional on a holdout occurring.

indicates that the average negotiation took place in a group of 3.2 members; however, this varies considerably, with 10 percent having five or more concurrent negotiations. As mentioned, there is a dramatic difference in the number of concurrent negotiations between paper mills and all other industries, with a mean of 14.6 in paper, and 1.4 outside it (most negotiations occur ''on their own'' outside the pulp and paper mills sector). Despite this, as we shall see, there is ample variation in the number of concurrent negotiations both inside and outside paper from which to identify the effects of information spillovers in bargaining and test our model's implications.

V. Empirical Analysis

A. *Raw Correlations*

Some raw empirical correlations relevant to the three main hypotheses examined in this paper are presented in Tables 3 through 5, respectively. Table 3 focuses on the relation between the number of concurrent negotiations and the incidence and duration of holdouts. Strikingly, when all industries are considered together, both holdout incidence and duration are monotonically increasing with the number of concurrent negotiations within the same three-digit industry, as predicted by a ''waiting game'' model of holdouts. Overall, mean holdout duration ranges from about 70 days when there is only one negotiation expiring at a time, to 126 days when there are 11 or more. When paper and other industries are considered separately, a similar monotonic pattern is observed for holdout duration outside paper, and for holdout incidence in the pulp and paper industry, though the other two series are noisier. Bearing in mind that these figures do not control for a number of potentially important determinants of holdouts, including for example industry effects outside of the paper industry and considerable observable heterogeneity

TABLE 4—MEAN NUMBER OF DAYS BETWEEN SUCCESSIVE HOLDOUT EXPIRATIONS, BY RANK IN NEGOTIATION GROUP

	Days between holdouts ranked:							
	0 and 1	1 and 2	2 and 3	3 and 4	4 and 5	5 and 6	6 and 7	7 and 8
(a) Whole sample	45.1 (3.5)	35.4 (3.3)	27.8 (5.0)	22.0 (5.7)	28.8 (7.2)	12.7 (3.9)	16.0 (9.1)	8.0 (3.6)
(b) Excluding paper	42.3 (3.7)	34.4 (3.6)	27.6 (5.7)	33.3 (9.6)	41.1 (12.4)	21.0 (10.5)	—[a] —	—[a] —
(c) Paper only	57.8 (9.6)	38.1 (8.7)	28.1 (9.2)	7.3 (2.2)	20.8 (8.6)	10.8 (4.2)	16.0 (9.1)	8.0 (3.6)

Notes: Table excludes negotiation groups with only one member. Standard errors in parentheses.
[a] No observations.

other dimensions within pulp and paper mills, we see them as providing some prima facie support for our "waiting game" model of holdouts.[19]

Table 4 focuses on the "bunching" of holdout expirations, or more precisely on the idea that holdouts should be more likely to end once other holdouts in the industry have ended. It examines this issue by comparing the mean duration of the shortest holdout in the group to the mean elapsed time between the end of the first and second holdout, second and third, and so on.[20] (If the hazard for holdouts were constant and independent across bargaining pairs within a group these durations should all be the same.) Again, a very suggestive pattern of shrinking delay between successive holdouts is found in the aggregate data: while the mean length of the first holdout in a group is about 45 days, the gap between the fifth and sixth, or sixth and seventh shrinks to about 14 days. As before, the trends are noisier when the sample is broken down between paper and all other industries, especially outside paper where relatively few higher-order holdouts are observed. This pattern of declining "marginal" durations could, of course, be generated by either an increasing hazard rate for individual holdouts or nonindependence between the respective holdout durations (both of which are consistent with our model). In what follows, however, we will estimate hazard models which isolate the latter effect from the former.

Table 5 considers the question of whether, because information is better for "followers" than "leaders," fewer strikes occur among those holdouts which end later in a negotiating group.[21] Although the differences are relatively small, it shows that strikes are indeed more common outcomes of those holdouts which end first in a group, as expected from the information-spillovers model of bargaining set out in this paper.

B. *Hazard Models of Holdout Durations*

Of course, while suggestive, the raw correlations in Tables 3–5 could be caused by a number of factors other than those suggested by the information-spillover model of holdouts, including, for example, differences in the

---

[19] Strictly speaking, our model predicts there should be no holdouts when only one bargaining pair is negotiating at a time. The fact that we apparently observe holdouts in such cases could be due, in part, to unions' waiting for information from sources not measured here, including contracts with different expiration dates, in other industries, or even publicly released statistical information. More generally, of course, factors other than information externalities could also be causing holdouts; our intention in this paper is not to argue that information externalities are the only cause of holdouts, only an important one.

[20] The sample in Table 4 restricts attention to negotiating groups with more than one member. Ties (holdouts ending on the same calendar day) are ranked randomly. Holdouts of zero duration are included.

[21] The relative rarity of strikes forces us to combine all "followers" into a single group here, rather than considering higher-order ranks separately, as in the previous table.

TABLE 5—FRACTION OF HOLDOUTS ENDING IN STRIKES: "FIRST" HOLDOUTS VERSUS OTHERS IN A NEGOTIATING GROUP

	"First" holdouts	Others
(a) Whole sample	0.220	0.171
	(0.023)	(0.017)
(b) Excluding paper	0.220	0.142
	(0.026)	(0.022)
(c) Paper only	0.224	0.198
	(0.051)	(0.024)

Notes: Table excludes negotiation groups with only one member. Exact binomial standard errors in parentheses.

institutions of bargaining across industries, business cycle conditions, and various characteristics of the bargaining pairs themselves. To control for these factors, and to obtain more precise estimates of the effects of the key variables under study here, we conduct two regression analyses. The first is a Cox proportional hazard model of holdout durations, which is used to assess the impact of the number of bargaining pairs on holdout durations, as well as the bunching of holdout durations. In the following subsection we conduct a probit analysis of *how* holdouts end, to assess the effect of being a "follower" on the probability a given holdout ends in a strike.

Table 6 presents the coefficient estimates from a Cox proportional hazard model of holdout durations.[22] As is well known, in this model the hazard for ending the holdout in contract negotiation i at time t is given by

$$(10) \qquad H_i(t) = e^{\beta' X_{it}} H_0(t),$$

where $H_0(t)$ is the baseline hazard at time t, X_{it} is a (potentially time-dependent) vector of explanatory variables for contract negotiation i, and $\boldsymbol{\beta}$ is a vector of parameters to be estimated. As is also well known, this specification implies that the explanatory variables influence the scale of the hazard rate, but not the form of its dependence on time. In the Cox partial likelihood approach to estimation which we adopt here, the baseline hazard $H_0(t)$ is allowed to be fully flexible, but is not estimated directly.

Aside from the number of concurrent negotiations in each negotiation's industry, observed covariates for which we are able to control in Table 6 include the length of the contract being negotiated, the monthly index of industrial production, the logarithm of the number of workers in the bargaining unit, 11 dummies for months and 25 dummies for years during which previous contracts expired, and dummies for regions and industries.[23] Contract length is included because our model predicts holdouts should be longer when a longer contract is being negotiated. The monthly index of industrial production (deseasonalized and detrended) is designed to capture any cyclical factors affecting holdout durations. This variable is constructed as the residuals from a regression of a 12-month moving average of the logarithm of the index of industrial production on a cubic spline in time [see Harrison and Mark Stewart (1989) for details]. The size of a bargaining unit is included because of its documented impact on strikes (Harrison and Stewart, 1993) and the possibility of a similar effect on holdout duration as well. Finally, our model for individual holdout durations also includes a time-varying covariate equal to the number of *other* holdouts that have already ended in the negotiation's "negotiating group." The coefficient on this variable should capture whether—controlling for a number of other

---

[22] Because it uses only the information on the ranking of durations, the Cox partial likelihood model used here (unlike, for example, Weibull duration models) encounters no special problems with zero durations. It does, however, impose the same (proportional) effects of the coefficients on the hazard at zero as at all other durations. As informal tests of this restriction, we estimated probit models of holdout incidence, and two-stage Heckit models of log durations. The probit incidence models gave coefficient estimates which are broadly similar in sign and significance to the duration results presented here, though generally less significant; the Heckit models gave very similar results, in size and significance, to OLS models of log durations. We also estimated the current Cox model on the sample of positive holdouts only, with very similar results to those reported here.

[23] To preserve degrees of freedom, in the regression for pulp and paper mills only, the year dummies were replaced by a fifth-order polynomial in time. Month dummies were dropped as well. This is because a large majority of the "synchronized" contracts expired in the same month, making it impossible to distinguish the effects of negotiating synchronously from seasonal effects. The group size coefficients for pulp and paper should be interpreted accordingly.

TABLE 6—DETERMINANTS OF HOLDOUT DURATIONS: COX PROPORTIONAL HAZARD COEFFICIENTS

	All industries		All industries excluding paper		Paper only
	(1)	(2)	(3)	(4)	(5)
Number of concurrent negotiations	−0.088 (−12.97)	−0.058 (−6.89)	−0.310 (−8.22)	−0.226 (−5.11)	−0.035 (−3.59)
Agreement duration	−0.015 (−6.00)	−0.023 (−8.39)	−0.015 (−5.53)	−0.024 (−7.97)	−0.030 (−3.30)
Log of number of employees	0.132 (4.39)	0.099 (2.82)	0.143 (4.51)	0.128 (3.35)	0.242 (2.23)
Index of industrial production	−4.32 (−3.64)	−3.39 (−2.75)	−2.64 (−2.09)	−2.34 (−1.79)	−1.84 (−1.43)
Number of other holdouts already ended	0.127 (13.03)	0.113 (11.64)	0.618 (11.63)	0.575 (10.59)	0.082 (6.12)
Industry dummies	no	yes	no	yes	no
Log likelihood	−18724.79	−18601.39	−15930.82	−15825.4	−1784.81
Sample size	2747	2747	2375	2375	372

Notes: All regressions, except those for paper mills only, include 4 dummies for 5 regions (Ontario, Quebec, BC, Maritime, and Prairies), 11 dummies for months, and 25 dummies for years during which previous contracts expired. "Paper only" regressions exclude month dummies and use a fifth-order polynomial in time in place of year dummies. Asymptotic *t*-statistics are in parentheses. Number of concurrent negotiations are measured at the beginning of the negotiation period.

possible determinants of holdouts, as well as for a common, but flexible, baseline hazard structure across all negotiations—the hazard rate for holdout expirations rises after another holdout has ended in its negotiating group.

In all samples and specifications considered, the coefficient estimates in Table 6 strongly support the implication of our model that the number of concurrent negotiations is positively related to expected holdout durations. (More negotiations reduce the log hazard, which in turn raises durations.)[24] Not surprisingly, the estimated effect of negotiating group size tends to be smaller and less significant when industry dummies are included—indicating that some of the identification comes from the cross-industry variation in bargaining structure—but remains highly significant in all specifications.[25] Somewhat more surprisingly, the effect of group size is considerably larger in magnitude outside the pulp and paper industry than inside it: loosely speaking, outside the paper industry, adding one member to the negotiation group reduces the chances of a holdout ending at any instant by 20–30 percent. In pulp and paper mills the effect is about $3^1/_2$ percent, with an average effect when all industries are combined of 6–9 percent. We conclude that: (a) there is a positive effect of negotiation group size on holdout durations, as predicted by our model; and (b) this effect is however much smaller in pulp and paper mills than in other industries, likely because of a number of peculiarities associated with that industry, including the much larger negotiation group size observed there.[26]

---

[24] In some other specifications (not reported) we disaggregated the number of concurrent negotiations into those involving the same union, and those involving another union. This was done to account for the possibility that cooperating locals of the same union might develop their own mechanisms (for example, a rotating arrangement) to solve the information externality problem. The coefficients on the number of negotiations in the same and other unions were, however, remarkably similar, suggesting that cooperation among locals in the timing of negotiations may not be an important factor in determining holdout durations.

[25] Essentially, the specifications with industry dummies identify the effects of group size by comparing unions whose previous contracts expire synchronously with other unions in their industry to unions in the same industry who negotiate asynchronously.

The coefficient on the duration of contracts in all columns of Table 6 is always negative and statistically significant, indicating that longer contracts give rise to longer holdouts, as is predicted by Proposition 4. The similarly robust positive and statistically significant coefficient on the size of a bargaining unit suggests that holdouts are shorter in larger bargaining units. The coefficient on the index of industrial production is negative but generally insignificant, providing some weak support for the notion that, unlike strike durations, holdout durations are procyclical (Harrison and Stewart, 1989). Finally, the positive and highly significant coefficient on the number of other holdouts already ended in the industry provides very strong evidence that holdouts are more likely to end once another holdout has ended in the industry, as is predicted by the "information spillover" hypothesis and suggested by Table 4. Again, the estimated magnitudes of these effects are much smaller in the paper industry than elsewhere: having one more holdout end in your industry raises the (instantaneous) probability of your holdout ending by about 60 percent outside paper, and by only about 8 percent in pulp and paper mills. Table 6 of course reinforces the results of Table 4 by showing that the apparent dependence of higher-order holdouts on earlier ones persists when we control for a number of covariates including industry fixed effects, and when we allow for a fully flexible baseline hazard common to all holdouts.

C. *How Holdouts End: Probit Models*

To determine whether holdouts which end later are less likely to end in strikes, as predicted by the information externalities in our model, we estimated a probit model in which the dependent variable equals 1 if a holdout ends in a strike and equals 0 if a holdout ends in a settlement.[27] In our analysis, the effect of negotiating later in a group is measured by a "follower" dummy variable, which distinguishes those holdouts which are first to end in a negotiating group from others. Other covariates are the same as were used in Table 6.

Coefficient estimates from the "strike" probits are reported in Table 7. According to Table 7, unions appear more willing to go on strike when negotiating longer agreements, and when their agreements involve a larger number of workers. Business conditions do not appear to play a significant role in strike incidence. Further, as expected from our model, holdouts which end later are less likely to end in a strike, at least outside the paper industry, although the effects are not as significant as those for the hypotheses examined in Table 6. Finally, and interestingly, there is also some weak evidence that, controlling for whether one is a "follower" in negotiations or not, larger negotiating groups are more prone to strikes. This is an issue on which our model is silent (the only effect of bargaining unit structure on strikes in our model is one which works *through* the presence of "followers" with access to better information), and potentially worthy of further exploration.

VI. Conclusion

In this paper we propose a new explanation of the most common form of labor contract disputes: holdouts. Unlike previous models, our model focuses on the information externality which arises among unions and firms in the same industry. Holdouts are modeled as a delaying tactic, employed by unions to obtain information

---

[26] One potential explanation of the smaller marginal effect of group size in the paper industry is of course nonlinearity in the effects of group size itself [recall, from (9), that our model predicts a diminishing marginal effect of group size on the hazard]. When we added group size squared to the specifications in Table 6, however, it was never significant. In some cases adding the squared term even made the linear term insignificant, even though joint tests of the linear and squared terms always indicated a very high degree of significance. We conclude that there is insufficient information in the current data to make accurate inferences about the functional form of the group size-hazard relationship.

[27] An alternative way to address this question would be to estimate a competing hazard model of holdouts, with strikes and settlements as alternative destinations, then simulate out the expected relative fractions ending in strikes after different histories in the negotiating group, together with their standard errors. As all of the holdout durations in our data are complete, this approach—a principal advantage of which is its ability to handle incomplete spells—seems to be an excessively cumbersome, and not obviously superior method of addressing the question at hand.

TABLE 7—HOW HOLDOUTS END: PROBIT MODELS OF THE PROBABILITY OF ENDING IN A STRIKE

	All industries		All industries excluding paper		Paper only
	(1)	(2)	(3)	(4)	(5)
Number of concurrent negotiations	0.016 (2.04)	0.023 (2.08)	0.040 (0.96)	0.106 (2.08)	0.002 (0.13)
Agreement duration	0.035 (8.90)	0.030 (7.20)	0.032 (7.71)	0.027 (5.87)	0.078 (4.84)
Log of number of employees	0.197 (4.84)	0.118 (2.44)	0.236 (5.48)	0.169 (3.29)	−0.327 (−1.83)
Index of industrial production	0.675 (0.34)	1.26 (0.61)	0.507 (0.24)	0.895 (0.41)	1.90 (0.88)
"Follower" dummy[a]	−0.297 (−2.80)	−0.182 (−1.61)	−0.441 (−3.06)	−0.341 (−2.42)	0.060 (0.13)
Industry dummies	no	yes	no	yes	no
Log likelihood	−1299.7	−1232.2	−1135.8	−1070.2	−153.4
Sample size	2747	2747	2375	2375	362[b]

Notes: Dependent variable = 1 if a holdout ends in a strike; = 0 if a holdout ends in a settlement. All regressions include 4 dummies for 5 regions. All regressions except those for paper only (column 5) include 11 dummies for months, and 25 dummies for years in which the previous contract expired. Paper only regressions include a fifth-order polynomial in time, and no month dummies. Asymptotic *t*-statistics are in parentheses.

[a] The "follower" dummy equals one if a holdout has ended within the current negotiation's group, and zero otherwise.

[b] Ten observations in prairie region had to be dropped because they predict failure perfectly.

from bargaining outcomes of other unions in their industry. A main implication of the model, on which previous models of holdouts are silent, is the following: as the number of negotiations in the model expands, unions have a greater incentive to delay their wage settlements in order to learn from the outcomes of other negotiations, generating longer holdouts. We test this implication using a large sample of contract negotiations in Canadian manufacturing industries from 1965–1988, with encouraging results. Other predictions of the model which are confirmed by the data are: (i) holdout durations are "bunched" within industries (once one holdout has ended, others are more likely to end also); and (ii) holdouts which end later in a "negotiating group" are less likely to end in strikes.

While we view our empirical results in this paper more as suggestive than as conclusive evidence in favor of the model presented here, and while we see that model as complementary rather than competing with existing models of holdouts (e.g., Cramton and Tracy, 1992), we do see an important lesson emerging from the model's consistency with the evidence. In particular, we believe that these results strongly suggest that future work on wage negotiations and strikes can gain by paying closer attention to the process of information spillovers across bargaining pairs, and more broadly to the context of other negotiations in the industry. Unfortunately, despite its key role in the industrial relations literature, and despite its potential importance in explaining both intranational and cross-national differences in the performance of industrial relations systems, this context has been largely neglected in economists' analyses of strikes and other bargaining disputes to date.

APPENDIX A: PROOFS

PROOF OF LEMMA 1:

1. This results immediately from the fact that each union's offer, once made, is a take-it-or-leave-it offer.

2. If the two unions make wage demands simultaneously, neither union has a previous negotiation from which to acquire any information. The optimal wage demand depends on

the value of each union's prior about its firm's state. Specifically, union i makes a "high" wage demand of π_g if $g_i > b$, and makes a low wage demand π_b if $g_i < b$.

3. If union i makes a wage demand before union j, union i's optimal wage demand will be exactly the same as in point 2 above: it has no previous negotiation from which to acquire any information. Union j's optimal wage demand, however, now depends on what it observes from the preceding negotiations, since this affects its optimally updated prior, p_j, that firm j is in the good state. Specifically, union j demands π_g if $p_j > b$ and π_b if $p_j < b$. Union j's posterior in each of the three possible outcomes of the preceding negotiations is calculated as follows:

(i) *Low wage settlement* (π_b) *in union i*. When $g_i < b$, union i makes a low wage demand of π_b, which firm i will always accept. Union j clearly learns nothing from the outcome of the previous negotiations. Its posterior is the same as its prior $p_j = g_j$.

(ii) *High wage settlement* (π_g) *without a strike in union i*. When $g_i > b$, union i makes a high wage demand of π_g. If firm i accepts this demand, it must be in the good state. Union j's updated prior that firm j is in the good state is then $p_j = P(jG|iG)$.

(iii) *Strike in union i*. When $g_i > b$, union i makes a high wage demand of π_g. If firm i rejects this wage demand and a strike ensues, it must be in the bad state. Union j's updated prior that firm j is in the good state is then $p_j = P(jG|iB)$.

PROOF OF PROPOSITION 1:

Suppose that union i ($i = 1, 2$) makes a wage demand at time $t \in [0, T]$. If union i makes a wage demand before or at the same time as the other union j, it makes a low wage demand of π_b, since its prior is $p < b$; if union i makes a wage demand after the other union j, it observes a low wage settlement of π_b in union j. It clearly learns nothing from the outcomes of previous negotiations. Its updated prior is $p < b$ and its optimal wage demand is then π_b. Therefore union i will always make a low wage demand of π_b, no matter when the other union makes a wage demand. Union i's expected payoff is:

$$\text{(A1)} \qquad \int_0^t w_i e^{-rt}\,dt + \int_t^T \pi_b e^{-rt}\,dt.$$

It is a dominant strategy for union i to make a low wage demand of π_b at time 0, since $w_i < \pi_b$ by assumption.

PROOF OF PROPOSITION 2:

Suppose union 1 makes a wage demand at time $t \in [0, T]$. By Lemma 1, union 1 makes a high wage demand of π_g no matter when union 2 makes a demand. Union 1's expected payoff is

$$\text{(A2)} \qquad \int_0^t w_1 e^{-rt}\,dt + \int_t^T g_1\pi_g e^{-rt}\,dt.$$

Union 1 maximizes its expected payoff by making a high wage demand of π_g at time 0 since $g_1\pi_g > w_1$.

Given union 1 makes a high demand of π_g at time 0, union 2 chooses a time to make a demand in order to maximize its expected payoff. If union 2 chooses time 0 to make a wage demand, it demands π_b and gets the expected payoff

$$\text{(A3)} \qquad \int_0^T \pi_b e^{-rt}\,dt.$$

If union 2 chooses time $t > 0$ to make a demand, i.e., makes a demand after union 1, union 2's optimal wage demand depends upon what it observes in the preceding negotiations. If union 1 strikes, firm 1 must be in the bad state, union 2's updated prior that firm 2 is in the good state is then $p_2 = P(2G|1B) < g_2 < b$. Its optimal strategy is of course simply to demand π_b; if union 1 does not strike, firm 1 must be in the good state, and union 2's updated prior is then $p_2 = P(2G|1G)$. Union 2 makes a high wage demand of π_g if and only if $P(2G|1G) > b$.

To summarize, when union 2 chooses $t > 0$ to make a wage demand, its expected payoff is

$$\text{(A4)} \qquad \int_0^t w_2 e^{-rt}\,dt + \int_t^T \pi_b e^{-rt}\,dt$$

if $P(2G|1G) < b$, and it is

$$\text{(A5)} \qquad \int_0^t w_2 e^{-rt}\,dt + \int_t^T \{\, g_1 P(2G|1G)\pi_g + (1 - g_1)\pi_b \,\}\, e^{-rt}\,dt$$

if $P(2G|1G) > b$.

Comparing union 2's expected payoffs (A3), (A4), and (A5), we have the proof.

PROOF OF PROPOSITION 3:

1. If union i is the first to make a wage demand, it makes a high wage demand of π_g; if union i is the second to make a demand, it makes a high wage demand of π_g if there was no strike in the preceding negotiation between union j and firm j, and makes a low wage demand of π_b if there was a strike. This follows directly from union i's updated prior p_i: if there was a strike in the preceding negotiation, its updated prior that firm i is in the good state is $p_i = P(iG|jG) > g_i > b$; otherwise if there was no strike, its updated prior is $p_i = P(iG|jB) = 1 - P(iB|jB) < b$.

Suppose that in a mixed strategy equilibrium, union i makes a wage demand with probability $h_i(t)dt$ between t and $t + dt$, conditional on neither having made a demand. In equilibrium, each union must be indifferent between making a demand and not making a demand at time t. Conditional on neither union having made a wage demand at time t, if union i makes the optimal high wage demand of π_g at t, its expected payoff is

$$\text{(A6)}\qquad \int_t^T g_i\pi_g e^{-rt}\,dt = g_i\pi_g\,\frac{e^{-rt}-e^{-rT}}{r}.$$

If union i waits until time $t + dt$, it is paid the wage w_i. However, with probability $h_j(t)dt$, the other union j makes a wage demand during this short interval of time. In this case, union i makes an optimal demand right after union j and from then on gets

$$\text{(A7)}\qquad \int_{t+dt}^T [g_jP(iG|jG)\pi_g + (1-g_j)\pi_b]\,e^{-rt}\,dt$$
$$= [g_jP(iG|jG)\pi_g + (1-g_j)\pi_b] \times \frac{e^{-r(t+dt)}-e^{-rT}}{r}.$$

If union j does not make a demand between t and $t + dt$, union i (by indifference) should be willing to make an optimal wage demand π_g and from time $t + dt$ on it now gets

$$\text{(A8)}\qquad \int_{t+dt}^T g_i\pi_g e^{-rt}\,dt = g_i\pi_g\,\frac{e^{-r(t+dt)}-e^{-rT}}{r}.$$

Union i is indifferent between making a wage demand and waiting until $t + dt$ if and only if

$$\text{(A9)}\qquad g_i\pi_g\,\frac{e^{-rt}-e^{-rT}}{r}$$
$$= e^{-rt}w_i\,dt$$
$$+ h_j(t)dt[g_jP(iG|jG)\pi_g + (1-g_j)\pi_b] \times \frac{e^{-r(t+dt)}-e^{-rT}}{r}$$
$$+ (1-h_j(t)dt)\,g_i\pi_g\,\frac{e^{-r(t+dt)}-e^{-rT}}{r}.$$

Rearranging the above the equation,

$$\text{(A10)}\qquad (g_i\pi_g - w_i)e^{-rt}dt$$
$$= h_j(t)dt\{g_jP(iG|jG)\pi_g + (1-g_j)\pi_b - g_i\pi_g\} \times \frac{e^{-r(t+dt)}-e^{-rT}}{r}.$$

The right-hand side of the equation is the benefit from the opportunity to learn from others, and the left-hand side is the cost associated with the delayed higher wage settlement. A union trades off the benefit against the cost to determine when to make a wage demand.

Solving (A10) as $dt \to 0$, we have

$$\text{(A11)}\qquad h_j(t) = \frac{r[g_i\pi_g - w_i]}{(1-g_j)\pi_g[b-P(iG|jB)]}\cdot\frac{1}{1-e^{-r(T-t)}},\quad \text{for } j = 1, 2.$$

2. Suppose union i ($i = 1, 2$) makes a wage demand at time t. When the firms' profits are highly uncorrelated such that $P(1B|2B) < 1 - b$ and $P(2B|1B) < 1 - b$, union i makes a high wage demand of π_g no matter when the other union j makes a demand and what the outcomes in the preceding negotiations are. To see this, note that when union i makes a wage demand before or at the same time as union j, it makes a high wage demand of π_g since its prior $g_i > b$. When union i makes a demand after union j made the optimal wage demand π_g, its updated prior is $p_i = P(iG|jG) > g_i > b$ if firm j accepts and it is $p_i = P(iG|jB) = 1 - P(iB|jB) > b$ if firm j rejects. Thus union i's updated prior is greater than b no matter what the outcomes in the preceding negotiation are and its optimal strategy is of course simply to make a high wage demand of π_g.

When union i chooses time t to make a wage demand, it will always demand π_g from the above argument and thus gets

$$\text{(A12)} \quad \int_0^t w_i e^{-rt}\,dt + \int_t^T g_i \pi_g e^{-rt}\,dt.$$

It chooses time 0 to make the high wage demand of π_g to maximize its expected payoff. Similarly we can show that union j will also make a high wage demand of π_g immediately at time 0.

3. By a similar argument as in point 2 above, in equilibrium union 1 will always make a high wage demand of π_g at time 0. Given union 1's strategy, if union 2 chooses time 0 to make a demand, it will demand π_g and get

$$\text{(A13)} \quad \int_0^T g_2 \pi_g e^{-rt}\,dt.$$

If union 2 chooses time $t > 0$ to make a demand and union 1 does not strike, its updated prior is $p_2 = P(2G|1G) > b$ and its optimal strategy is to demand π_g. If union 2 moves at $t > 0$ and union 1 strikes, its updated prior is $p_2 = P(2G|1B) < b$ and its optimal strategy is to demand π_b. Union 2 thus expects to get

$$\text{(A14)} \quad \int_0^t w_2 e^{-rt}\,dt + \int_t^T [g_1 P(2G|1G)\pi_g + (1 - g_1)\pi_b]\, e^{-rt}\,dt.$$

From (A13) and (A14), we can see that uni 2 will choose to make a wage demand rig after union 1 to maximize its payoff.

PROOF OF PROPOSITION 4:

1. At time t, the hazard that holdou end is $h_1(t) + h_2(t)$. From Corollary 1, w have

$$\text{(A15)} \quad \partial[h_1(t) + h_2(t)]/\partial w_1 = \partial[h_1(t) + h_2(t)]/\partial w_2 = -\frac{r}{(1-p)\pi_b} \cdot \frac{1}{1 - e^{-r(T-t)}} < 0$$

Therefore, higher wage settlements lead t lower hazards and longer holdouts.

2. It is easy to verify that

$$\text{(A16)} \quad \partial[h_1(t) + h_2(t)]/\partial p = \frac{r(2p\pi_g - w_1 - w_2)}{(1-p)^2 \pi_b} \cdot \frac{1}{1 - e^{-r(T-}}$$

$$> 0.$$

A higher p leads to a higher hazard rate a thus a shorter holdout.

3. Differentiating with respect to T yields

$$\text{(A17)} \quad \partial[h_1(t) + h_2(t)]/\partial T = \frac{r(2p\pi_g - w_1 - w_2)}{(1 - p)\pi_b} \times \frac{(-r)e^{-r(T-t)}}{[1 - e^{-r(T-t)}]^2} < 0.$$

A longer contract duration thus leads to longer holdout.

PROOF OF PROPOSITION 5:

1. This follows immediately from a similar argument to that in the proof of Proposition 1.

2. The proof is strictly parallel to part 1 of Proposition 3, based on indifference between making an offer at times t and $t + dt$. In particular, the payoff to making an offer immediately is given by equation (A6), substituting p for g_i (the firms now have the same probability of being in the good state). The payoff to waiting, if *any one* of the other unions makes a wage demand in the interval $t + dt$, is given by (A7), substituting (again because of perfect correlation) $[p\pi_g + (1 - p)\pi_b]$ for $[g_j P(iG \mid jG)\pi_g + (1 - g_j)\pi_b]$. The payoff to waiting if no other union makes an offer in the interval is given by (A8), again substituting p for g_i. The probability of any other union making an offer in the interval $t + dt$ is $(n - 1)h(t)dt$. Indifference between making a demand today and waiting implies:

$$\text{(A20)} \qquad h(t) = \frac{r(p\pi_g - w)}{(n-1)(1-p)\pi_b} \cdot \frac{1}{1 - e^{-r(T-t)}}.$$

APPENDIX B: DATA

TABLE B1—SUMMARY STATISTICS: THREE-DIGIT MANUFACTURING INDUSTRIES

SIC code, industry	Number of observations	Incidence (percent)		Mean duration (days)		Mean unit size[c]	Mean contract length[d]	Mean negotiating goup size
		Strike[a]	Holdout[b]	Strike	Holdout			
101 Meat/Poultry	76	10.1	77.6	49.5	87.1	1812	25.4	3.5
102 Fish Products	47	25.5	95.7	14.6	108.0	1902	22.8	1.4
103 Fruit/Vegetables Process	13	15.4	92.3	21.5	69.1	1011	25.1	1.0
104 Dairy Products	21	4.8	95.2	13.0	83.0	933	24.1	1.1
105 Flour/Cereal	21	4.8	95.2	10.0	87.1	536	23.0	1.0
107 Bakery Products	60	16.7	88.3	36.1	73.9	622	25.0	1.2
108 Miscellaneous Food	80	12.5	76.3	29.1	80.6	637	24.2	1.2
109 Beverages	71	15.5	91.6	36.0	126.8	811	30.5	3.2
153 Tobacco Products	48	4.2	79.2	19.0	80.8	905	23.9	1.4
162 Rubber Products	83	25.3	90.4	42.9	54.3	872	34.7	1.2
174 Shoe	18	11.1	94.4	16.5	60.0	693	23.6	1.0
181 Cotton Yarn	53	15.1	84.9	43.3	78.9	1005	28.2	1.4
183 Man-made Fibre	62	24.2	77.4	32.6	82.4	945	26.3	1.0
188 Automotive Fabric Accessories	13	23.1	53.9	24.7	27.6	710	32.4	1.0
189 Miscellaneous Textiles	11	45.5	81.8	24.0	61.0	523	26.2	1.0
243 Men's Clothing	87	5.8	79.3	14.2	66.2	1723	25.8	1.6
244 Women's Clothing	29	6.9	82.8	6.0	30.1	3060	32.7	1.9
246 Fur Goods	15	20.0	66.7	4.3	51.3	1003	28.7	1.5
251 Sawmills	50	38.0	70.0	38.6	76.5	6878	23.9	2.4
252 Veneer/Plywood	10	20.0	80.0	34.5	93.5	567	22.3	1.0
261 Household Furniture	13	38.5	76.9	18.0	44.9	567	21.3	1.0
271 Pulp/Paper Mills	372	19.4	94.4	58.4	126.8	1092	28.4	14.6
273 Paper Box/Bag	10	10.0	100.0	8.0	140.8	801	27.9	1.0
274 Miscellaneous Paper	18	11.1	72.2	66.5	94.2	523	24.1	1.1
286 Commercial Printing	24	4.2	41.7	4.0	69.7	644	20.3	1.0
287 Platemaking	49	2.0	100.0	22.0	93.8	1093	23.6	1.4
289 Publishing/Printing	29	20.7	96.6	47.2	110.0	1020	25.9	1.1

TABLE B1—*Continued.*

SIC code, industry	Number of observations	Incidence (percent) Strike[a]	Incidence (percent) Holdout[b]	Mean duration (days) Strike	Mean duration (days) Holdout	Mean unit size[c]	Mean contract length[d]	Mean negoti group
291 Iron/Steel Mills	103	27.2	70.9	28.9	65.7	2214	30.2	1.8
292 Steel Pipe/Tube	31	16.1	71.0	43.0	77.7	815	33.5	1.2
295 Smelting/Refining	93	30.1	71.0	43.2	41.5	1736	28.1	1.2
296 Aluminum Rolling	17	29.4	82.3	88.6	52.5	888	30.9	1.0
297 Copper/Alloy Rolling	10	10.0	60.0	96.0	45.0	730	30.9	1.0
301 Boiler/Plate Works	18	16.7	77.8	39.0	61.4	670	25.8	1.0
302 Fabricated Metal	13	15.4	76.9	62.5	36.5	700	26.7	1.0
304 Metal Stamping	30	10.0	53.3	18.7	70.0	1114	34.7	1.1
311 Agricultural Implement	39	43.6	74.4	15.2	34.3	1774	28.1	1.0
315 Miscellaneous Machinery	76	18.4	72.4	33.9	68.0	802	26.4	1.1
321 Aircraft/Parts	93	29.0	77.4	56.2	77.1	1448	28.0	1.1
323 Motor Vehicle	78	44.9	80.8	26.5	49.5	4955	31.7	1.7
324 Truck Body/Trailer	11	27.3	90.9	17.3	53.3	635	29.5	1.0
325 Motor Vehicle Parts	89	34.8	65.2	26.9	37.3	820	34.1	1.0
326 Rail Rolling Stock	35	31.4	77.1	60.9	76.9	917	28.5	1.0
327 Shipbuilding/Repair	84	21.4	85.7	35.4	103.4	997	28.0	1.5
332 Major Appliances	65	23.1	92.3	32.5	45.6	836	28.7	1.2
334 Radio/TV	37	8.1	75.7	26.3	46.9	957	21.5	1.0
335 Communications Equipment	130	19.2	79.2	25.3	75.8	1166	26.8	1.1
336 Electrical Industrial Equipment	34	41.2	88.2	28.2	58.5	2680	26.2	1.0
338 Electric Wire/Cable	14	14.3	85.7	26.0	34.0	648	21.9	1.0
339 Miscellaneous Electrical	24	25.0	75.0	56.2	73.9	599	27.5	1.0
356 Glass/Products	80	23.8	82.5	40.9	45.9	718	28.3	1.1
359 Miscellaneous Nonmetallic	17	29.4	94.1	24.0	59.4	524	23.5	1.0
378 Industrial Chemicals	64	15.6	65.6	46.6	51.9	860	21.7	1.1
379 Miscellaneous Chemical	37	29.7	51.4	32.1	145.9	639	24.3	1.0
391 Scientific Equipment	23	21.7	73.9	27.6	36.7	823	23.7	1.0
393 Sporting Goods/Toys	19	10.5	57.9	25.5	64.6	652	26.1	1.0
All Manufacturing	2747	21.0	80.8	37.4	79.6	1339	27.5	3.2

[a] Strike: strike occurs before contract is signed.
[b] Holdout: holdout occurs before contract is signed.
[c] Mean unit size: = mean number of workers in a bargaining unit.
[d] Mean duration = mean duration of the contract being negotiated, in months.

REFERENCES

Anderson, John C. "The Structure of Collective Bargaining," in John C. Anderson, Morley Gunderson, and Allen M. Ponak, eds., *Union-management relations in Canada*, 2nd Ed. Don Mills, Ontario: Addison-Wesley, 1989, pp. 209–34.

Banerjee, Abhijit V. "A Simple Model of Herd Behavior." *Quarterly Journal of Economics*, August 1992, *107*(3), pp. 797–818.

Card, David. "Longitudinal Analysis of Str Activity." *Journal of Labor Economic* April 1988, *6*(2), pp. 147–76.

______. "Strikes and Wages: A Test of Asymmetric-Information Model." *Qu terly Journal of Economics*, August 199 *105*(3), pp. 625–59.

Chamley, Christophe and Gale, Douglas. "Inf mation Revelation and Strategic Delay i Model of Investment." *Econometrica*, Se tember 1994, *62*(5), pp. 1065–85.

Cramton, Peter C. and Tracy, Joseph S. "Strikes and Holdouts in Wage Bargaining: Theory and Data." *American Economic Review*, March 1992, *82*(1), pp. 100–21.

———. "The Determinants of U.S. Labor Disputes." *Journal of Labor Economics*, April 1994, *12*(2), pp. 180–209.

Fernandez, Raquel and Glazer, Jacob. "Striking for a Bargain Between Two Completely Informed Agents." *American Economic Review*, March 1991, *81*(1), pp. 240–52.

Fudenberg, Drew and Tirole, Jean. *Game theory*. Cambridge, MA: MIT Press, 1991.

Gu, Wulong. "Adoption and Diffusion of Technological Innovations: Informational Externalities and Strategic Delays." Mimeo, University of Saskatchewan, 1996.

Gunderson, Morley; Kervin, John and Reid, Frank. "The Effect of Labour Relations Legislation on Strike Incidence." *Canadian Journal of Economics*, November 1989, *22*(4), pp. 779–94.

Harrison, Alan. "Strikes and Contracts in Postwar Canada." Mimeo, McMaster University, 1994.

Harrison, Alan and Stewart, Mark. "Cyclical Fluctuations in Strike Durations." *American Economic Review*, September 1989, *79*(4), pp. 827–41.

———. "Strike Duration and Strike Size." Working Paper No. 93-05, McMaster University, 1993.

Hayes, Beth. "Unions and Strikes with Asymmetric Information." *Journal of Labor Economics*, January 1984, *2*(1), pp. 57–83.

Kuhn, Peter and Gu, Wulong. "The Benefits of Learning From Others: Sequential Wage Bargaining." Mimeo, McMaster University, 1996.

———. "Centralization and Strikes." *Labour Economics*, 1998 (forthcoming).

McFadden, Daniel L. and Train, Kenneth E. "Consumers' Evaluation of New Products: Learning from Self and Others." *Journal of Political Economy*, August 1996, *104*(4), pp. 683–703.

Osborne, Martin J. "The Role of Risk Aversion in a Simple Bargaining Model," in Alvin E. Roth, ed., *Game theoretical models of bargaining*. Cambridge: Cambridge University Press, 1985, pp. 181–213.

Rob, Rafael. "Learning and Capacity Expansion Under Demand Uncertainty." *Review of Economic Studies*, July 1991, *58*(4), pp. 655–75.

Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size

By JOSHUA D. ANGRIST AND WILLIAM N. EVANS*

Research on the labor-supply consequences of childbearing is complicated by the endogeneity of fertility. This study uses parental preferences for a mixed sibling-sex composition to construct instrumental variables (IV) estimates of the effect of childbearing on labor supply. IV estimates for women are significant but smaller than ordinary least-squares estimates. The IV are also smaller for more educated women and show no impact of family size on husbands' labor supply. A comparison of estimates using sibling-sex composition and twins instruments implies that the impact of a third child disappears when the child reaches age 13. (*JEL* J13, J22)

An understanding of the relationship between fertility and labor supply is important for a number of theoretical and practical reasons. First, economists and demographers have developed a variety of models linking the family and the labor market. Empirical studies of childbearing and labor supply are sometimes seen as tests of these models (e.g., Reuben Gronau, 1973; Mark R. Rosenzweig and Kenneth I. Wolpin, 1980b; T. Paul Schultz, 1990). Second, the link between fertility and labor supply might partly explain the postwar increase in women's labor-force participation rates if having fewer children causes an increase in labor-force attachment (Mary T. Coleman and John Pencavel, 1993). Evidence for this thesis includes Claudia Goldin's (1995) study, which shows that few women in the 1940's and 1950's birth cohorts were able to combine childbearing with strong labor-force attachment. Other researchers have also drawn a link between fertility-induced withdrawals from the labor force and lower wages of women (e.g., Gronau, 1988; Sanders Korenman and David Neumark, 1992). So perhaps childbearing keeps women from developing their careers.

Any success in disentangling the causal mechanisms linking fertility and labor supply should shed light on other substantive issues as well. For example, reductions in female labor supply could increase the total time parents devote to child care, making at least some children better off (see, e.g., Frank P. Stafford, 1987; Francine Blau and Adam J. Grossberg, 1992). Some theories of family behavior also suggest that changes in wives' earnings affect marital stability (Becker et al., 1977; Becker, 1985).

Not surprisingly, given the wide and longstanding interest in the connection between childbearing and labor supply, hundreds of empirical studies report estimates of this relationship. The vast majority of these studies find a negative correlation between fertility (or family size) and female labor supply.[1] A noted in two recent literature surveys, however, the interpretation of these correlations remains unclear. In his assessment of the "economics of the family," Robert J. Willis (1987 p. 74) writes, "... it has proven difficult to find enough well-measured exogenous variables to permit cause and effect relationships to be extracted from correlations among factors such as the delay of marriage, decline of childbearing, growth of divorce, and increased female labor force participation ... ." Martin Browning (1992 p. 1435) expresses similar views: "... although we have a number of robust correlations, there are very few credible inferences that can be drawn from them."[2]

---

* Angrist: Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139; Evans: Department of Economics, University of Maryland, College Park, MD 20742. The authors thank seminar participants at the Council of Economic Advisors, Brown University, the University of Maryland, MIT, the University of Michigan, Northwestern University, Princeton University, the University of Virginia, and the July 1996 NBER Labor Studies meetings, as well as Susan Athey, Peter Diamond, Jon Gruber, Judy Hellerstein, Jim Poterba, David Weil, and three anonymous referees for helpful discussions or comments. John Johnson and Amanda Honeycutt provided excellent research assistance. Special thanks go to Duncan Thomas who stimulated our interest in the subject of parental sex preferences. The authors bear sole responsibility for the content of this paper.

[1] There is less work on the effects of children on husbands' labor supply. See Pencavel (1986 Table 1.17) for a few estimates, which suggest a positive association between fathers' labor supply and the number of children. The relationship between husbands' and wives' labor s

Skepticism regarding the causal interpretation of associations between fertility and labor supply arises in part from the fact that there are strong theoretical reasons to believe that fertility and labor supply are jointly determined (see, e.g., Schultz, 1981, or Goldin, 1990). In fact, this endogeneity is reflected in the academic research agenda. On one hand, papers on labor supply often treat child-status variables as regressors in hours of work equations, while on the other hand, economic demographers and others discuss regressions and models that are meant to characterize the impact of wages or measures of labor-force attachment on fertility. Since fertility variables cannot be both dependent and exogenous at the same time, it seems unlikely that either sort of regression has a causal interpretation.[3]

This paper focuses on the causal link running from fertility to the work effort of both men and women. Our main contribution is the use of a new instrumental variables (IV) strategy based on the sibling sex mix in families with two or more children. This instrument exploits the widely observed phenomenon of parental preferences for a mixed sibling-sex composition. In particular, parents of same-sex siblings are significantly and substantially more likely to go on to have an additional child.[4] Because sex mix is virtually randomly assigned, a dummy for whether the sex of the second child matches the sex of the first child provides a plausible instrument for further childbearing among women with at least two children. Moreover, in spite of the fact that the sibling sex mix is obviously a function of the sex of both children, we present a range of evidence which suggests that there is little possibility that any secular impact of the sex of offspring contaminates the IV estimates.

We also compare results generated using sex mix as an instrument to results generated using twins to construct instruments. Twinning at first birth has been used in a number of previous studies to estimate the relationship between

---

ply is discussed by, among others, Orley Ashenfelter and James J. Heckman (1974), Heckman and Thomas E. MaCurdy (1980), and Jonathan Gruber and Julie Berry Cullen (1996).

[2] The survey by Alice Nakamura and Masao Nakamura (1992) argues that a search for exogenous variation is so difficult it is not even fruitful (pp. 60–61).

[3] In the chapter on models of marital status and childbearing in the *Handbook of Labor Economics*, Mark R. Montgomery and James Trussel (1986 p. 205) note:

> "One of the rites of passage for a labor economist involves the estimation of a Probit model for female labor force participation. It is standard practice for the Probit equation to include some indicators for a woman's marital status and the number and age distribution of her children. In estimating such a model, the labor economist veers dangerously close to a theory of household formation, childbearing, and labor supply; namely, that household formation and fertility can be safely taken as exogenous with respect to a woman's supply of hours."

Many of the papers cited in the *Handbook* chapter on female labor supply (Mark R. Killingsworth and Heckman, 1986) fit this description. A widely cited paper that discusses regressions of fertility measures on measures of earnings is Jacob Mincer (1963). A more recent example is Heckman and James R. Walker (1990), who use nonlinear techniques. See also Schultz (1981 p. 171) who asks: "What is cause and what is effect?" in a discussion of William P. Butz and Michael P. Ward (1979), and other demographic studies of the relationship between fertility and labor-force participation.

[4] Charles F. Westoff et al. (1963) were among the first to report preferences for a mix. In a survey of desired fertility and a follow-up study of actual fertility among couples with two children, they found that parents of two boys or two girls both desired and ultimately had more children than parents of mixed pairs. See Nancy E. Williamson (1976) for an international review. After completing the first draft of this paper in July 1996, we learned of concurrent work using sex-preference instruments to estimate the effect of fertility on female labor supply in the United Kingdom (Maria Iacovou, 1996).

childbearing and labor-market outcomes. Examples include Stephen G. Bronars and Jeff Grogger (1994) and Jaisri Gangadharan and Joshua L. Rosenbloom (1996), both of which use large Census samples, and Rozenzweig and Wolpin (1980a, b). Here, our interest is primarily in the comparison between twins estimates and results using the same-sex instruments. We therefore focus on multiple second births, so that both twinning and the sex-mix instrument identify the impact of moving from the second to the third child. Because third-born children who are twins are older than other third-born children, the juxtaposition of estimates based on twinning and the sex mix allows us to compare the effects of fertility on labor supply when the children are different ages. By combining the twins and same-sex instruments, we can estimate the time it takes for the labor-supply consequences of childbearing to disappear.

Section I discusses the data and the sex-mix instruments' first stage, and briefly describes how sex mix can be incorporated into standard economic models of fertility. Section II presents the main set of empirical results on fertility and labor supply, including an analysis of effects in subgroups defined by husbands' earnings and mothers' schooling. Section III compares the estimates using sex-mix instruments to estimates based on twins. Section IV discusses the empirical findings in light of recent trends in female labor-force participation, and Section V concludes.

I. Data, Descriptive Statistics, and First-Stage Relationships

A. *Data and Descriptive Statistics*

The sex-mix estimation strategy is implemented using information on labor supply, the sex of mothers' first two children, and an indicator of multiple births in the 1980 and 1990 Census Public Use Micro Samples (PUMS). To motivate the empirical work, Table 1 reports labor-force participation rates and the probability of additional childbearing among women aged 21–35 and aged 36–50 in the 1970, 1980, and 1990 PUMS. Data for 1970 are from the 1/100 state file (U.S. Department of Commerce, Bureau of the Census, 1972); data for 1980 and 1990 are from the 5-percent samples (U.S. Department of Commerce, Bureau of the Census, 1983, 1995). The table shows substantial declines in fertility and increases in labor supply in both age-groups. Statistics for women aged 21–35 with at least two children show a similar pattern, as do the statistics for women aged 36–50.

There is no retrospective fertility information in the PUMS data sets other than the number of children ever born. We therefore matched children to mothers within households in a manner similar to that described in the appendix to Angrist and Evans (1996a). Briefly, we attached people in a household labeled as "child" in the primary relationship code to a female householder or the spouse of a male householder. In households with multiple families, detailed relationship codes as well as subfamily identifiers were used to pair children with mothers. We deleted any mothers for whom the number of children in the household did not match the reported number of children ever born.[5] Using the sex of the oldest two children, we defined same-sex sibling pairs in both Censuses.

Because the Census does not track children across households, the sample is limited to mothers aged 21–35 whose oldest child was less than 18 years of age at the time of the Census. Few women younger than age 21 have two children, while a child over age 17 is increasingly likely to have moved to a different household. Restricting the women's age-group to less than or equal to 35 means the age 18 cutoff for firstborn children does not generate a highly selected sample. Data from the Fertility, Birth Expectations, and Marital History Supplement to the June 1990 Current Population Survey (CPS) show that among women aged 35 with two or more children, at least 9 percent have an oldest child younger than age 18. This fraction falls to 85 percent at age 3 but is equal to 100 percent for women aged 3 or younger. Although women aged 21–3 with at least two children may appear to constitute an unusually young high-fertility group

---

[5] Note also that the sample is restricted to women for whom the reported values of age and sex of their two oldest children were not allocated by the U.S. Bureau of the Census.

TABLE 1—FERTILITY AND LABOR-SUPPLY MEASURES

Sample	1970 PUMS	1980 PUMS	1990 PUMS
Women aged 21–35			
Mean children ever born	1.78	1.27	1.18
Percent with 2 or more children	52.10	40.40	37.60
Percent worked last year	60.00	73.40	79.30
Observations	203,918	1,326,631	1,478,546
Women aged 36–50			
Mean children ever born	2.85	2.86	2.15
Percent with 2 or more children	73.40	78.50	68.90
Percent worked last year	57.30	66.70	78.50
Observations	181,502	852,204	1,253,095
Women aged 21–35 with 2 or more children			
Mean children ever born	3.06	2.61	2.57
Percent with more than 2 children	55.60	39.90	39.10
Percent worked last year	44.80	58.00	66.60
Observations	106,239	535,587	577,397
Married women aged 21–35 with 2 or more children			
Mean children ever born	3.02	2.58	2.53
Percent with more than 2 children	54.90	39.00	37.50
Percent worked last year	41.80	55.80	67.50
Observations	91,286	436,483	439,408

Notes: The 1970 PUMS data are from the 1/100 state file. The 1980 and 1990 data are from the 5-percent PUMS. Calculations from the 1990 PUMS use sample weights. The married samples include women married at the time of the Census.

Our tabulations of the June 1990 CPS show that over half of all women aged 28–35 fall into this group. The proportion is lower for women aged 21–27 but still includes at least one-quarter of the entire age cohort.[6]

The empirical analysis is conducted on two subsamples from each Census data set. The first includes all women with two or more children. The second includes only married women because this is the sample that many economic theories of household production (e.g., Gronau, 1973) are meant to describe. The married sample is also used to explore the impact of children on fathers' labor supply.[7] The 1980 married sample is restricted to couples who were married at the time of the Census, married only once, and married at the time of their first birth. There are 394,835 observations in the full 1980 sample and 254,654 observations in the 1980 married sample (64 percent of the total). Information on the timing of first marriage and the number of marriages is not available in the 1990 PUMS, so that the 1990 married sample includes all women who were married at the time of the Census. The full 1990 sample includes 380,007 women and the married 1990 sample includes 301,588 women (79 percent of the total—higher than for 1980 because the 1990 sample-selection rule is less restrictive).

---

[6] It is also worth noting that a substantial fraction of the change in completed family size between 1970 and 1990 occurred at parities greater than 2. Our tabulations of 1970 and 1990 Census data show that about 71 percent of women aged 40–55 in both years had two or more children. The proportion having three or more children, however, fell from about 0.47 in 1970 to 0.39 in 1990.

Descriptive statistics and variable definitions for covariates, instruments, and dependent variables are given in Table 2. The covariate of primary interest in our labor-supply models is the indicator *More than 2 children.* The first instrumental variable for *More than 2 children* is the indicator *Same sex.* The table also shows averages for the two components of *Same sex,* the indicators *Two boys* and *Two girls.* Among all women with two children in 1980, 40.2 percent had a third child. The corresponding figure for 1990 is 37.5. In both samples, just over 50 percent of all two-child families had children of the same sex and just over 51 percent of first births were boys.

Labor-supply estimates are also computed using multiple second births to generate instruments. In the 1980 PUMS, multiple births are defined as siblings having the same age and quarter of birth. The mean for this indicator of twin births, which we call *Twins-2,* is 0.0085 in the 1980 full sample and 0.0083 in the 1980 married sample.[8] For purposes of comparison, we drew a sample of all second births born to women aged 21–35 from the 1976 Vital Statistics Natality Data tapes (National Center for Health Statistics). This data set contains a 5 percent sample of all births in the country and should provide an accurate estimate of twinning probabilities for the women in our sample. Data from 1976 offer a useful comparison since roughly 40 percent of second children in our Census sample were born 1976–197 The vital statistics data imply a second-birth twinning probability of 0.0079, just slightly lower than the probability we estimate using 1980 Census data.

Quarter of birth is not reported in the 1990 PUMS, so multiple births in 1990 were defined as children reported to be of the same age. Using this procedure, we calculated that 1.2 percent of all second births in 1990 were multiple births. This naturally produces a much higher estimate of the number of twins since two children born in the same 12-month period are classified as twins. We used data from the 1980 PUMS to estimate the error in twin rates calculated using age in years only. Using age in years to define twins in the 1980 data generates an estimated twin rate of 0.01185, which is 35 percent larger than the value we calculate using age in years and quarter of birth. We therefore restricted the analysis using twins to data from the 1980 PUMS.

Demographic and labor-supply variables described in the lower half of Table 2, include measures of mother's age, age at first birth, years of education, and indicators for race and ethnic background. We also report values for the husbands of women in the married sample. The labor-supply variables are based on Census questions concerning work in 1979 or 1989. These variables measure whether respondents *Worked for pay,* their *Weeks worked,* usual *Hours/week,* and annual *Lab*

---

[7] One reason estimates are presented for the full sample as well as for the married sample is that conditioning on marital status raises the possibility that selection bias affects estimates in the selected sample.

[8] Of the 3,356 multiple second births in the all-wo sample, only 23 were triplets or higher-plurality bir Therefore, we use the terms multiple births and interchangeably.

TABLE 2—DESCRIPTIVE STATISTICS, WOMEN AGED 21–35 WITH 2 OR MORE CHILDREN

	Means and (standard deviations)					
	1980 PUMS			1990 PUMS		
	All women	Married couples		All women	Married couples	
Variable		Wives	Husbands		Wives	Husbands
Children ever born	2.55 (0.81)	2.51 (0.77)	—	2.50 (0.76)	2.48 (0.74)	—
More than 2 children (=1 if mother had more than 2 children, =0 otherwise)	0.402 (0.490)	0.381 (0.486)	—	0.375 (0.484)	0.367 (0.482)	—
Boy 1st (s_1) (=1 if first child was a boy)	0.511 (0.500)	0.514 (0.500)	—	0.512 (0.500)	0.514 (0.500)	—
Boy 2nd (s_2) (=1 if second child was a boy)	0.511 (0.500)	0.513 (0.500)	—	0.511 (0.500)	0.512 (0.500)	—
Two boys (=1 if first two children were boys)	0.264 (0.441)	0.266 (0.442)	—	0.264 (0.441)	0.265 (0.441)	—
Two girls (=1 if first two children were girls)	0.242 (0.428)	0.239 (0.427)	—	0.241 (0.428)	0.239 (0.426)	—
Same sex (=1 if first two children were the same sex)	0.506 (0.500)	0.506 (0.500)	—	0.505 (0.500)	0.503 (0.500)	—
Twins-2 (=1 if second birth was a twin)	0.0085 (0.0920)	0.0083 (0.0908)	—	0.012 (0.108)	0.011 (0.105)	—
Age	30.1 (3.5)	30.4 (3.4)	33.0 (4.6)	30.4 (3.5)	30.7 (3.3)	33.4 (4.8)
Age at first birth (parent's age in years when first child was born)	20.1 (2.9)	20.8 (2.9)	24.0 (4.0)	21.8 (3.5)	22.4 (3.5)	25.1 (4.7)
Worked for pay (=1 if worked for pay in year prior to census)	0.565 (0.496)	0.528 (0.499)	0.977 (0.150)	0.662 (0.473)	0.667 (0.471)	0.968 (0.175)
Weeks worked (weeks worked in year prior to census)	20.8 (22.3)	19.0 (21.9)	48.0 (10.5)	26.2 (22.9)	26.4 (22.9)	47.1 (12.0)
Hours/week (average hours worked per week)	18.8 (18.9)	16.7 (18.3)	43.5 (12.3)	22.5 (19.1)	22.2 (18.9)	44.0 (13.3)
Labor income (labor earnings in year prior to census, in 1995 dollars)	7,160 (10,804)	6,250 (10,211)	38,919 (25,014)	9,550 (13,071)	9,616 (13,238)	36,623 (30,283)
Family income (family income in year prior to census, in 1995 dollars)	42,342 (26,563)	47,646 (25,821)	—	42,558 (34,692)	49,196 (34,740)	—
Non-wife income (family income minus wife's labor income, in 1995 dollars)	—	41,635 (24,734)	—	—	39,580 (31,892)	—
Number of observations	394,835	254,654	254,654	380,007	301,588	301,588

Notes: The samples include women aged 21–35 with two or more children except for women whose second child is less than a year old. In the 1980 PUMS, the married women sample refers to women who were married at the time of their first birth, married at the time of the survey, and married once. In the 1990 PUMS, the married women are those married at the time of the Census.

income. The latter three variables are set to zero for those who did not work for pay during the year. The final two variables in the table are measures of *Family income* and, for the married sample, a variable called *Non-wife income* computed as family income minus the wife's labor income.[9] The descriptive statistics show that women's labor-force participation rates, weeks and hours worked, and age at first birth increased between 1980 and 1990. Women's real (1995 dollar) earnings increased substantially as well, especially for married women, while real *Non-wife income* declined. It should also be noted that the variances of earnings, husband's earnings, and family income increased substantially over this period.

Finally, note that husband's age at first birth was calculated assuming that the husband is the father of the children in the household. All of the husbands' variables are computed based in this assumption, which seems plausible for the 1980 data since women in the 1980 married women sample were married only once and they were married at the time of their first birth. Only 4.7 percent of the husbands in this sample were married before, and very few children live with their fathers after divorce. On the other hand, the 1990 match is probably not as good as the 1980 match. We therefore confirmed the basic first-stage relationships used in this paper with June CPS data, which includes true retrospective fertility information. See the appendix to our earlier paper (Angrist and Evans, 1996a) using a Census household match for more on data problems and issues related to the match.

B. *Sex Mix and Fertility*

The phenomenon of parental preferences for a mixed sibling-sex composition has been documented in a number of studies. For example, Yoram Ben-Porath and Finis Welch (1976) found that in the 1970 Census, 56 percent of families with either two boys or two girls had a third birth, whereas only 51 percent of families with one boy and one girl had a third child.

The theoretical impact of sex mix on fertility can be captured in the standard quantity-quality model of fertility, originally outlined by Becker and Gregg H. Lewis (1973) and Becker and Nigel Tomes (1976), and extended in detail by Rosenzweig and Wolpin (1980a). In these models, parents derive utility from the number of children and a complementary good, "child quality," which enters the utility function and budget constraint in proportion to the number of children. Child quality is generated through the purchase of inputs and the expenditure of parents' time in home production. Ben-Porath and Welch (1980) describe the sex mix as something that determines child quality in quantity/quality models. Alternately, the impact of sex preferences can be modeled using state-dependent utility. Suppose a mother already has $n_x \geq 1$ children and she is trying to decide how many additional children to have (n_c). If parents prefer a mixed sibling-sex composition, then the same-sex sibling composition reduces the utility from n_x. This in turn raises the marginal utility of n_c, increasing the chances that parents will try to have additional children. Twinning can similarly be incorporated into the model as a shock to n_x that cannot be fully offset by future fertility choices. For a more detailed theoretical discussion, see our working paper (Angrist and Evans, 1996b).

Table 3 reports estimates of the impact of child sex and the sex mix on fertility similar to those in Ben-Porath and Welch (1976). The first panel looks at sex preferences in families with one or more children by showing the fraction of women with at least one child who had a second child, conditional on the sex of the first child. The third row of this panel shows the difference by sex. In spite of the fact that attitudinal surveys suggest many couples would prefer more boys than girls, or prefer their firstborn child to be male (see, e.g., Williamson, 1976), there is only one subsample (all women in the 1990 PUMS) where subsequent fertility is a function of the sex of the first child. Even in this case, the impact of the sex of the firstborn on fertility is very small.

The second panel of Table 3 documents the relationship between the fraction of women who have a third child and the sex of the first two children. The first three rows from the

---

[9] In the few cases where there were negative or zero family-income values, we set the variables equal to one when computing logs. Family income and person wage and salary income are top coded at $75,000 in the 1980 Census. In the 1990 Census, family income is top coded at $999,999 and individual wage and salary income is top coded at $140,000, with state medians of top-coded values substituted for the top code.

TABLE 3—FRACTION OF FAMILIES THAT HAD ANOTHER CHILD BY PARITY AND SEX OF CHILDREN

	All women				Married women			
	1980 PUMS (649,887 observations)		1990 PUMS (627,362 observations)		1980 PUMS (410,333 observations)		1990 PUMS (477,798 observations)	
Sex of first child in families with one or more children	Fraction of sample	Fraction that had another child	Fraction of sample	Fraction that had another child	Fraction of sample	Fraction that had another child	Fraction of sample	Fraction that had another child
(1) one girl	0.488	0.694 (0.001)	0.489	0.665 (0.001)	0.485	0.720 (0.001)	0.487	0.698 (0.001)
(2) one boy	0.512	0.694 (0.001)	0.511	0.667 (0.001)	0.515	0.720 (0.001)	0.513	0.699 (0.001)
difference (2) − (1)	—	0.000 (0.001)	—	0.002 (0.001)	—	0.000 (0.001)	—	0.001 (0.001)

	All women				Married women			
	1980 PUMS (394,835 observations)		1990 PUMS (380,007 observations)		1980 PUMS (254,654 observations)		1990 PUMS (301,588 observations)	
Sex of first two children in families with two or more children	Fraction of sample	Fraction that had another child	Fraction of sample	Fraction that had another child	Fraction of sample	Fraction that had another child	Fraction of sample	Fraction that had another child
one boy, one girl	0.494	0.372 (0.001)	0.495	0.344 (0.001)	0.494	0.346 (0.001)	0.497	0.331 (0.001)
two girls	0.242	0.441 (0.002)	0.241	0.412 (0.002)	0.239	0.425 (0.002)	0.239	0.408 (0.002)
two boys	0.264	0.423 (0.002)	0.264	0.401 (0.002)	0.266	0.404 (0.002)	0.264	0.396 (0.002)
(1) one boy, one girl	0.494	0.372 (0.001)	0.495	0.344 (0.001)	0.494	0.346 (0.001)	0.497	0.331 (0.001)
(2) both same sex	0.506	0.432 (0.001)	0.505	0.407 (0.001)	0.506	0.414 (0.001)	0.503	0.401 (0.001)
difference (2) − (1)	—	0.060 (0.002)	—	0.063 (0.002)	—	0.068 (0.002)	—	0.070 (0.002)

Notes: The samples are the same as in Table 2. Standard errors are reported in parentheses.

section show the sample characteristics of women in the following groups: those with one boy and one girl, those with two girls, and those with two boys. The next two rows report estimates for women with two children of the same sex and for women with one boy and one girl. The final row reports the differences between the same-sex and mixed-sex group averages.

Both data sets and samples suggest that women with two children of the same sex are much more likely to have a third child than the mothers of one boy and one girl. For example, in the 1980 all-women sample, only 37.2 percent of women with one boy and one girl have a third child, compared to 43.2 for women with two girls or two boys. The relationship between sex mix and the probability of additional childbearing is even larger for married women, reaching a precisely estimated 7-percentage-point difference in the 1990 Census. This is approximately 21 percent of the rate of additional childbearing among women with one boy and one girl. Finally, we note that the relationship between sex mix and childbearing is confirmed in data from the fertility supplements to the June 1980, 1985, and 1990 CPS. This is important because, unlike the Census where information about children is partly based on our household match, the

June CPS contains detailed fertility histories for each woman, including information on the dates of birth and sex of each child.

The virtual random assignment of *Same sex* makes it very likely that reduced-form regressions of fertility and labor-supply outcomes on the instruments have a causal interpretation. One simple check on this claim is to compare the demographic characteristics of people who have same-sex and mixed-sex sibling compositions. Table 4 reports *Same sex* contrasts for mother's age, age at first birth, race, ethnicity, and years of education in the 1980 and 1990 all-women samples. Even in these large samples, none of the contrasts is significantly different from zero at the 5-percent level. The magnitude of the differences by *Same sex* is also very small. For example, the difference of −0.0028 for years of education in 1980 represents 0.02 percent of the sample mean years of schooling, which is about 13.

In contrast with the small and insignificant differences in demographic characteristics by *Same sex,* there are some large and precisely estimated differences in mean demographic variables by twin status. The estimates in the final column of Table 4 replicate the well-known result that twins are more likely for older women (John A. H. Waterhouse, 1950) and for blacks (Ntinos Myrianthopoulos, 1970). Women with more years of schooling are also more likely to have twins, although this probably reflects more childbearing at older ages among more educated women.

II. Fertility and Labor Supply

A. *Wald Estimates*

Because sibling-sex composition is virtually randomly assigned, simple statistical techniques can be used to illustrate how the sex-mix IV strategy identifies the effect of fertility on parents' labor supply. Consider the bivariate regression model,

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{1}$$

where y_i is a measure of labor supply and x_i is the endogenous fertility measure of interest. Let z_i denote the binary instrument, *Same sex*. The IV estimate of β in this equation is

$$\beta_{\mathrm{IV}} = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0), \tag{2}$$

where $\bar{y}_1$ is the mean of y_i for those observations with $z_i = 1$ and other terms are similarly defined. The numerator and denominator capture the reduced-form relationships between y_i and z_i and between x_i and z_i. The IV method attributes any effect of z_i on y_i to the effect of z_i on x_i.

Although equation (1) is written as a bivariate regression with constant coefficients, Guido W. Imbens and Angrist (1994) have shown that β_{IV} can be interpreted as a local average treatment effect specific to the instrument, z_i. In this case, β_{IV} estimates the average effect of x_i on y_i for individuals whose fertility has been affected by their children's sex mix. Similarly, when z_i is an indicator of multiple births at the second pregnancy, *Twins-2*, the IV estimates reflect the effect of children on labor supply for those who have had more children than they otherwise would have because of twinning. For this reason, the *Twins-2* and *Same sex* instruments do not necessarily identify the same average effect.

The first six columns of Table 5 report the components of β_{IV} when *Same sex* is used as the instrument in the all-women samples from the 1980 and 1990 PUMS. The last three columns report corresponding results from the 1980 PUMS using the *Twins-2* instrument. The first two rows of the table show the denominator of the Wald estimate, $\bar{x}_1 - \bar{x}_0$, for two possible choices of x_i. One is an indicator for having had a third child, *More than 2 children.* The other is total *Number of children.* The effect of the *Same sex* instrument on *More than 2 children,* equal to the difference in means reported at the bottom of Table 3, is 0.06 in 1980 and 0.063 in 1990. The effect of *Same sex* on *Number of children* is 0.077 in 1980 and 0.084 in 1990. The effect of *Twins-2* on the probability of having a third birth in 1980 is 0.60, and the effect of *Twins-2* on *Number of children* is 0.81.

Below the estimates of $\bar{x}_1 - \bar{x}_0$, columns (1) and (4) of Table 5 report $\bar{y}_1 - \bar{y}_0$ for alternative outcomes using the *Same sex* instrument. These results show that in addition

TABLE 4—DIFFERENCES IN MEANS FOR DEMOGRAPHIC VARIABLES BY *SAME SEX* AND *TWINS*-2

	Difference in means (standard error)		
	By *Same sex*		By *Twins-2*
Variable	1980 PUMS	1990 PUMS	1980 PUMS
Age	−0.0147 (0.0112)	0.0174 (0.0112)	0.2505 (0.0607)
Age at first birth	0.0162 (0.0094)	−0.0074 (0.0114)	0.2233 (0.0510)
Black	0.0003 (0.0010)	0.0021 (0.0011)	0.0300 (0.0056)
White	0.0003 (0.0012)	−0.0006 (0.0013)	−0.0210 (0.0066)
Other race	−0.0006 (0.0005)	−0.0014 (0.0009)	−0.0090 (0.0041)
Hispanic	−0.0014 (0.0009)	−0.0007 (0.0010)	−0.0069 (0.0047)
Years of education	−0.0028 (0.0076)	0.0100 (0.0074)	0.0940 (0.0415)

Notes: The samples are the same as in Table 2. Standard errors are reported in parentheses.

to having more children than women with one boy and one girl, women with two children of the same sex have a lower probability of working, work fewer weeks per year and fewer hours per week, and have lower annual earnings and lower family income. All but the final result is statistically significant in both Census years.

The Wald estimates for 1980 calculated by dividing $\bar{y}_1 - \bar{y}_0$ by $\bar{x}_1 - \bar{x}_0$ when x_i is *More than 2 children* imply that having more than two children reduced labor supply by 13.3 (−0.008/0.06) percentage points, weeks worked by about 6.4 weeks, hours of work per week by 5.2, and labor income by just over $2,200 per year. The results for 1990 are also negative, though (with the exception of family income) somewhat smaller. The Wald estimates calculated using the effect of *Same sex* on total *Number of children* put these effects in per-child terms. In per-child terms, the estimates are about 0.78 as large in 1980 and 0.75 as large in 1990 as the estimates produced with *More than 2 children* in the denominator.

The last three columns in the table show that women whose second pregnancy resulted in twins are also less likely to work. With the exception of the estimate for family income, which is not very precise, the Wald estimates generated by *Twins-2*, reported in column (6), are lower than the Wald estimates based on *Same sex*. In Section III, we explore the comparison between *Same sex* and *Twins-2* estimates further and show how they can be reconciled.

As with the *Same sex* estimates, *Twins-2* estimates in per-child terms are necessarily smaller than estimates treating the indicator *More than 2 children* as the endogenous regressor. But the factor of proportionality connecting the per-child and *More than 2 children* estimates using *Twins-2* is also 0.75. It therefore makes little difference which denominator is used because estimates based on *More than 2 children* can always be converted into per-child estimates by multiplying by 0.75. We chose to discuss estimates treating *More than 2 children* as the endogenous regressor in the re-

TABLE 5—WALD ESTIMATES OF LABOR-SUPPLY MODELS

	1980 PUMS			1990 PUMS			1980 PUMS		
		Wald estimate using as covariate:			Wald estimate using as covariate:			Wald estimate using as covariate:	
Variable	Mean difference by *Same sex*	*More than 2 children*	*Number of children*	Mean difference by *Same sex*	*More than 2 children*	*Number of children*	Mean difference by *Twins-2*	*More than 2 children*	*Number of children*
More than 2 children	0.0600 (0.0016)	—	—	0.0628 (0.0016)	—	—	0.6031 (0.0084)	—	—
Number of children	0.0765 (0.0026)	—	—	0.0836 (0.0025)	—	—	0.8094 (0.0139)	—	—
Worked for pay	−0.0080 (0.0016)	−0.133 (0.026)	−0.104 (0.021)	−0.0053 (0.0015)	−0.084 (0.024)	−0.063 (0.018)	−0.0459 (0.0086)	−0.076 (0.014)	−0.05 (0.01
Weeks worked	−0.3826 (0.0709)	−6.38 (1.17)	−5.00 (0.92)	−0.3233 (0.0743)	−5.15 (1.17)	−3.87 (0.88)	−1.982 (0.386)	−3.28 (0.63)	−2.4 (0.4
Hours/week	−0.3110 (0.0602)	−5.18 (1.00)	−4.07 (0.78)	−0.2363 (0.0620)	−3.76 (0.98)	−2.83 (0.73)	−1.979 (0.327)	−3.28 (0.54)	−2.4 (0.4
Labor income	−132.5 (34.4)	−2208.8 (569.2)	−1732.4 (446.3)	−119.4 (42.4)	−1901.4 (670.3)	−1428.0 (502.6)	−570.8 (186.9)	−946.4 (308.6)	−705 (229
ln(*Family income*)	−0.0018 (0.0041)	−0.029 (0.068)	−0.023 (0.054)	−0.0085 (0.0047)	−0.136 (0.074)	−0.102 (0.056)	−0.0341 (0.0223)	−0.057 (0.037)	−0.04 (0.02

Notes: The samples are the same as in Table 2. Standard errors are reported in parentheses.

mainder of the paper because this emphasizes the fact that the fertility increment induced by either instrument is a move from two to more than two children.

B. *Two-Stage Least-Squares Estimation*

While the Wald estimates provide a simple illustration of how the instruments identify the effect of children on labor supply, the rest of the paper discusses two-stage least-squares (2SLS) and ordinary least-squares (OLS) estimates of regression models relating labor-market outcomes to fertility and a variety of exogenous covariates. 2SLS estimation allows us to accomplish three things. First, even if there is no association between the instrument and exogenous covariates, as suggested by Table 4, controlling for exogenous covariates can lead to more precise estimates if the treatment effects are roughly constant across groups.

Second, we can use 2SLS to control for any secular additive effects of child sex when using *Same sex* as an instrument. This is desirable because *Same sex* is an interaction term involving the sex of the first two children, and therefore potentially correlated with the sex of either child. To see this, let s_1 and s_2 be indicators for male firstborn and second-born children. The instrument can be written as

$$(3)\qquad Same\ sex = s_1 s_2 + (1 - s_1)(1 - s_2).$$

Assuming that child sex is independent and identically distributed (i.i.d.) over children, the population regression of *Same sex* on either s_j produces a slope coefficient equal to $2E[s_j] - 1$, which is zero only if $E[s_j] = ½$. Since the probability of giving birth to a male child is 0.51, there is a slight positive association between *Same sex* and the sex of each child. This correlation is a concern only if

---

[10] Proof: Assuming child sex is i.i.d., we have $E[s_1] = E[s_2]$ and $E[s_1 s_2] = E[s_j]^2$. Therefore, $\mathrm{Cov}(Same\ sex, s_j) = E[s_j](E[s_j] - E[Same\ sex])$. Some manipulation gives $E[s_j] - E[Same\ sex] = (1 - E[s_j])(2E[s_j] - 1)$. Since the variance of s_j is $E[s_j](1 - E[s_j])$, the regression coefficient is $(2E[s_j] - 1)$.

effects labor supply for reasons other than family size. Such effects could arise if the sex of offspring affects the father's commitment to the family (see, e.g., Philip S. Morgan et al., 1988) or changes the way parents treat their children (Kristin F. Butcher and Anne Case, 1994; Duncan Thomas, 1994). Secular effects of sex mix on labor supply could also be generated by the fact that boys are more likely than girls to have disabilities (see, e.g., Angrist and Victor Lavy, 1996) since having a disabled child might change parents' behavior. Adding s_1 and s_2 as regressors to the estimating equations reduces the likelihood of omitted-variables bias from these sources.

Of course, controls for additive effects can only eliminate bias from omitted variables with effects that are additive in the number of children. However, a third advantage of the 2SLS framework is that it allows us to exploit the fact that the *Same sex* instrument can be decomposed into two instruments, leading to an overidentified model. In particular, the separate indicators, *Two boys* $[s_1 s_2]$ and *Two girls* $[(1 - s_1)(1 - s_2)]$, are both available as potential instruments. This is useful because bias from any secular effects of child sex on labor supply should be different for these two instruments, while the labor-supply consequences of childbearing seem likely to be independent of whether *Same sex* equals *Two boys* or *Two girls*. A natural specification test is therefore the conventional instrument-error overidentification test statistic for 2SLS estimation using both instruments, since this is the same as a test for whether the *Two boys* and *Two girls* instruments give the same estimate when used separately.[11]

The following regression models are used to link labor-supply variables for husbands and wives to the endogenous *More than 2* variable, x_i, and the list of exogenous covariates, including additive effects for the sex of each child:

$$(4) \qquad y_i = \alpha_0' \mathbf{w}_i + \alpha_1 s_{1i} + \alpha_2 s_{2i} + \beta x_i + \varepsilon_i,$$

where $\mathbf{w}_i$ is a vector of demographic variables, and s_{1i} and s_{2i} are indicators for the sex of the first two children of mother i. Initially, $\mathbf{w}_i$ is limited to variables that are clearly exogenous to fertility: mother's age and age at first birth, plus race and Hispanic indicators. In the just-identified model where *Same sex* is the only instrument, the first-stage equation relating *More than 2 children* to sex mix is

$$(5) \qquad x_i = \pi_0' \mathbf{w}_i + \pi_1 s_{1i} + \pi_2 s_{2i} + \gamma(Same\ sex_i) + \eta_i,$$

where γ is the first-stage effect of the instrument.

The alternative identification strategy uses the two components of *Same sex*—*Two boys* and *Two girls*—as instruments for *More than 2 children.* In this case, however, either s_{1i} or s_{2i} must be dropped from the list of covariates because s_{1i}, s_{2i}, $s_{1i}s_{2i}$, and $(1 - s_{1i})(1 - s_{2i})$ are linearly dependent. We chose to drop s_{2i} (the results are not sensitive to this choice, or to the elimination of both s_{1i} and s_{2i}, as we show below). In this case, the equation of interest becomes

$$(6) \qquad y_i = \alpha_0' \mathbf{w}_i + \alpha_1 s_{1i} + \beta_i x_i + \varepsilon_i.$$

The first-stage relationship between x_i and sex mix is

$$(7) \qquad x_i = \pi_0' \mathbf{w}_i + \pi_1 s_{1i} + \gamma_0(Two\ boys_i) + \gamma_1(Two\ girls_i) + \eta_i,$$

where $Two\ boys_i = s_{1i}s_{2i}$ and $Two\ girls_i = (1 - s_{1i})(1 - s_{2i})$.

C. 2SLS Results

The first-stage results linking sex mix and fertility are reported in Table 6. In the top half of the table, we report results from the 1980 PUMS. These estimates show that women in 1980 with same-sex children are estimated to be 6.2 percentage points more likely to have a third child in a model with covariates. The corresponding estimate for married women is 6.9 percent. The

---

[11] See Whitney K. Newey and Kenneth D. West (1987) for this interpretation of overidentification tests. Angrist (1991) discusses the dummy instrument case.

TABLE 6—OLS ESTIMATES OF *MORE THAN 2 CHILDREN* EQUATIONS

	All women			Married women		
Independent variable	(1)	(2)	(3)	(4)	(5)	(6)
1980 PUMS						
Boy 1st	—	−0.0080 (0.0015)	0.0001 (0.0021)	—	−0.0111 (0.0018)	−0.001 (0.00
Boy 2nd	—	−0.0081 (0.0015)	—	—	−0.0095 (0.0018)	—
Same sex	0.0600 (0.0016)	0.0617 (0.0015)	—	0.0675 (0.0019)	0.0694 (0.0018)	—
Two boys	—	—	0.0536 (0.0021)	—	—	0.059 (0.00
Two girls	—	—	0.0698 (0.0021)	—	—	0.07 (0.00
With other covariates	no	yes	yes	no	yes	yes
R^2	0.004	0.084	0.084	0.005	0.078	0.07
1990 PUMS						
Boy 1st	—	−0.0081 (0.0015)	−0.0083 (0.0022)	—	−0.0097 (0.0017)	−0.00 (0.002
Boy 2nd	—	0.0002 (0.0015)	—	—	−0.0011 (0.0017)	—
Same sex	0.0628 (0.0016)	(0.0623) (0.0015)	—	0.0702 (0.0018)	0.0703 (0.0017)	—
Two boys	—	—	0.0624 (0.0021)	—	—	0.069 (0.002
Two girls	—	—	0.0621 (0.0022)	—	—	0.071 (0.002
With other covariates	no	yes	yes	no	yes	yes
R^2	0.004	0.082	0.082	0.005	0.082	0.082

Notes: Other covariates in the models are indicators for *Age, Age at first birth, Black, Hispanic,* and *Other race.* Th variable *Boy 2nd* is excluded from columns (3) and (6). Standard errors are reported in parentheses.

t-statistics for these first-stage effects are well over 30. As in Table 3, the estimates for the 1990 PUMS (reported in the lower half of Table 6) are somewhat larger in both the full and married women samples.

Table 6 also provides some evidence of an association between having a male child and reduced childbearing at higher parities. Note, however, that the effect of *Boy 1st* in the 1980 data is explained entirely by the difference the effect of *Two boys* and *Two girls* wh these regressors are entered separately. other words, when the effects of sex mix allowed to differ by sex, there is no relatio ship between *Boy 1st* and fertility, although effect of *Same sex* on fertility in 1980 is larg for boys than for girls. The *Boy 1st* effects 1990 remain significant in all specificatio but they are very small, especially in comparison to the effects of the sex mix.

Next, we use the sex mix to estimate the effect of *More than 2 children* on measures of employment and earnings in 1980. Table 7 reports a set of OLS estimates and two sets of 2SLS estimates using *Same sex* and the pair of dummies *Two boys* and *Two girls* as instruments. The exogenous regressors are the same as in Table 6 (coefficients not reported). In models that use *Two boys* and *Two girls* as instruments, we dropped the *Boy 2nd* variable from the list of covariates. The first three columns show results for the full sample, the next three columns show results for married women, and the last three columns show results for the husbands of married women.

OLS estimates in both the full and married women sample suggest that the presence of a third child reduces the probability of working by about 17 percentage points, and causes weeks worked to fall by about 8–9 per year, hours per week to fall by 6–7, and family income to fall by about 13 percent. OLS estimates of earnings effects are $3,166 in the married sample and $3,768 in the full sample. Not surprisingly, all of these OLS estimates are very precisely estimated.

In contrast with the results for women, OLS estimates of the effect of *More than 2 children* on husbands' labor supply are small. Having a third child is estimated to reduce the probability a husband worked for pay by less than one percentage point. The impact of a third child on other measures of husbands' labor supply is also small, though precise enough to be significantly different from zero. The estimated effect on annual weeks worked is −0.90 and the estimate for hours per week is 0.25. The effect on husbands' earnings appears substantial (−$1,506), but this amount is still only about 3.9 percent of the average earnings of men in the sample.

The first set of 2SLS estimates uses *Same sex* alone as an instrument. In the full sample, the estimates (standard errors) for the dependent variables *Worked for pay, Weeks worked, Hours/week,* and *Labor income* models are −0.12 (0.025), −5.7 (1.1), −4.6 (0.95), and −1,961 (542). These results suggest that having a third child causes a 20–30-percent reduction in women's labor supply and earnings. One important finding is that estimates using *Same sex* as an instrument are smaller than the corresponding OLS estimates. This is true for the labor-supply estimates in the married women's sample as well, although here the gap between 2SLS and OLS estimates is not as large. Overall, however, the OLS estimates appear to exaggerate the causal effect of fertility on female labor supply.

It is also worth noting that the relationship between the OLS and 2SLS estimates is similar when the estimates are converted into per-child units. Above, we noted that 2SLS estimates treating *More than 2 children* as the endogenous regressor should be multiplied by about 0.75 to obtain estimates in per-child terms (i.e., with *Number of children* as the endogenous regressor). It turns out that the OLS estimates can be converted into per-child units by multiplying by about 0.7 using either 1980 or 1990 data. Thus, regardless of whether the estimates are cast in terms of the effect of having more than two children or in per-child units, the OLS estimates are considerably smaller than the 2SLS estimates.

In addition to differing from the OLS estimates, the estimates using *Same sex* as an instrument also differ from most of the 2SLS estimates previously reported in the literature on children and labor supply. In his review article, Browning (1992 p. 1469) notes that, "There is one salient difference between studies that take fertility as exogenous and those that take it as endogenous. In many of the latter it is found that fertility either has no effect on labor supply ... or it has a positive effect." Browning also points out that it is not clear from these estimates whether children really have no effect on female labor supply, or whether the instruments are too weak or simply poorly chosen. While the 2SLS estimates generated by *Same sex* are smaller than the corresponding OLS estimates, they are still negative, precise, and of a plausible magnitude.

In contrast to the female labor-supply estimates, there is little evidence of a relationship between having a third child and family income. Given the strong labor-supply effects, the weak impact on family income may seem surprising. There are a few potential explanations for this result. First, the lost income due to a reduction in mothers' labor supply could be made up by other family members. The small and statistically insignificant 2SLS estimates in the ln(*Non-wife income*) equations suggest that this is not the case. The most likely explanation is that the instrument is not powerful enough to detect the family-income consequences of childbearing. For example, in the married women sample, the third child reduces female earnings by about 21 percent and female labor income is (on average) 13 percent of total family income. If the third child does not alter husbands' labor supply, we would expect an estimated effect of *More than 2 children* in the ln(*Family income*) equations of roughly $-0.21 \times 0.13 = -0.027$, which is close to the reported estimate of -0.05. But the standard error for this estimate is slightly higher than 0.05, so that impacts this small cannot be precisely measured.

Gronau (1977 p. 1102) reports results suggesting that husbands increase their work effort in response to an increase in family size. Table 7 also reports estimates of the impact of the third child on husbands' labor supply in the 1980 married sample. While the OLS estimates show a small but significant (negative) relationship between husbands' labor supply and additional childbearing, estimates constructed using *Same sex* as an instrument generate no evidence of any effects on the labor supply of men. The standard errors on the 2SLS estimates for husbands' variables *Worked for pay, Weeks worked*, and *Hours/week* are actually smaller than the corresponding standard errors for women, and they are small enough so that modest positive or negative effects could be detected if they existed.

The labor-supply effects estimated using 1990 data are remarkably similar to those estimated for 1980. This can be seen in Table 8, which reports OLS estimates and 2SLS estimates for 1990 using *Same sex* and *Two boys* and *Two girls* as instruments. Some of the estimated effects are slightly smaller in 1990 than in 1980, but these differences are not statistically meaningful. One difference between the 1980 and 1990 results that does seem noteworthy is the larger negative impact of childbearing on married women's earnings in 1990, perhaps because of an increase in women's wages. This result may also be attributa the fact that married women are del childbearing (average age at first birt creased from 20.8 in 1980 to 22.4 in 199 this group), and therefore they have years of experience and higher wages they exit the workforce due to childbirth

Table 6 shows that mothers of two gir more likely than mothers of two boys to a third child. So the first-stage relationship fers for these two instruments. However, 2SLS estimates in Tables 7 and 8 show the additional predictive power provided separating the two components of *Same* does not change the coefficient estimates much or lead to an appreciable increase in cision using either the 1980 or 1990 data.

We noted above that the overidentifica test statistic associated with the use of *boys* and *Two girls* as instruments jointly for a difference between 2SLS estimates puted using only *Two boys* and 2SLS estim computed using only *Two girls*. The *p*-val for this test are reported in square bracket both Tables 7 and Table 8. The *p*-values the 1990 estimates suggest that it does not ter which instrument is used. In fact, the 2S estimates using *Two boys* and *Two girls* 1990 are remarkably close. On the other ha the 1980 2SLS estimates are consisten smaller when *Two girls* alone is used as instrument. Moreover, some of the *p*-val for estimates computed using 1980 data in cate a significant contrast between the *girls* and *Two boys* instruments, although marginal significance level is below 1 perc As in the 1990 data, however, in the 1980 both instruments are always associated more children and reduced labor supply.

D. *Other Specification Issues*

Other specification issues considered include the robustness of the results, the g erality of the results, and the validity of instruments. Because sex mix is essentia randomly assigned, the results reported Tables 7 and 8 are unchanged by altering basic set of covariates. For example, us data for married women from the 19 PUMS, we estimated models adding the lowing covariates to the vector $\mathbf{w}_i$: linear

TABLE 7—OLS AND 2SLS ESTIMATES OF LABOR-SUPPLY MODELS USING 1980 CENSUS DATA

	All women			Married women			Husbands of married women		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Estimation method	OLS	2SLS	2SLS	OLS	2SLS	2SLS	OLS	2SLS	2SLS
Instrument for *More than 2 children*	—	*Same sex*	*Two boys, Two girls*	—	*Same sex*	*Two boys, Two girls*	—	*Same sex*	*Two boys, Two girls*
Dependent variable:									
Worked for pay	−0.176 (0.002)	−0.120 (0.025)	−0.113 (0.025) [0.013]	−0.167 (0.002)	−0.120 (0.028)	−0.113 (0.028) [0.013]	−0.008 (0.001)	0.004 (0.009)	0.001 (0.008) [0.013]
Weeks worked	−8.97 (0.07)	−5.66 (1.11)	−5.37 (1.10) [0.017]	−8.05 (0.09)	−5.40 (1.20)	−5.16 (1.20) [0.071]	−0.82 (0.04)	0.59 (0.60)	0.45 (0.59) [0.030]
Hours/week	−6.66 (0.06)	−4.59 (0.95)	−4.37 (0.94) [0.030]	−6.02 (0.08)	−4.83 (1.02)	−4.61 (1.01) [0.049]	0.25 (0.05)	0.56 (0.70)	0.50 (0.69) [0.71]
Labor income	−3768.2 (35.4)	−1960.5 (541.5)	−1870.4 (538.5) [0.126]	−3165.7 (42.0)	−1344.8 (569.2)	−1321.2 (565.9) [0.703]	−1505.5 (103.5)	−1248.1 (1397.8)	−1382.3 (1388.9) (0.549)
ln(*Family income*)	−0.126 (0.004)	−0.038 (0.064)	−0.045 (0.064) [0.319]	−0.132 (0.004)	−0.051 (0.056)	−0.053 (0.056) [0.743]	—	—	—
ln(*Non-wife income*)	—	—	—	−0.053 (0.005)	0.023 (0.066)	0.016 (0.066) [0.297]	—	—	—

Notes: The table reports estimates of the coefficient on the *More than 2 children* variable in equations (4) and (6) in the text. Other covariates in the models are *Age, Age at first birth,* plus indicators for *Boy 1st, Boy 2nd, Black, Hispanic,* and *Other race.* The variable *Boy 2nd* is excluded from equation (6). The *p*-value for the test of overidentifying restrictions associated with equation (6) is shown in brackets. Standard errors are reported in parentheses.

quadratic terms in the wife's education, quadratic terms in wife's age, age at first birth, linear and quadratic terms in husband's age, husband's age at first birth and education, linear and quadratic terms in husband's labor income, and a full set of state dummy variables.[12] In these models, the 2SLS estimates (standard errors) of the *More than 2 children* coefficient have the following values: *Worked for pay,* −0.122 (0.027); *Weeks worked,* −5.45 (1.18); *Hours/week,* −5.04 (0.99); *Labor income,* −1,390 (555). All of these values are within 5 percent of the corresponding estimates from Table 7.

A referee and others who read earlier versions of this paper expressed concern about whether the results are likely to be representative of the impact of childbearing in general since the sample is restricted to women with two or more children and to women in a relatively young age-group. Estimates of the effect of going from two to more than two children do not necessarily generalize. On the other hand, we believe these results are likely to be of general interest because a significant fraction of the change in fertility between 1970 and 1990 was due to reductions in the number of women having more than two children. As noted in Section I, this fact is apparent in Census data on completed family size.

---

[12] Two of these covariates, years of education and husband's earnings, are potentially endogenous because they may be partly determined by fertility. For this reason, they were excluded from the main set of estimates.

TABLE 8—OLS AND 2SLS ESTIMATES OF LABOR-SUPPLY MODELS USING 1990 CENSUS DATA

	All women			Married women			Husbands of married women	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Estimation method	OLS	2SLS	2SLS	OLS	2SLS	2SLS	OLS	2SLS
Instrument for *More than 2 children*	—	*Same sex*	*Two boys, Two girls*	—	*Same sex*	*Two boys, Two girls*	—	*Same sex*
Dependent variable:								
Worked for pay	−0.155 (0.002)	−0.092 (0.024)	−0.092 (0.024) [0.743]	−0.147 (0.002)	−0.104 (0.024)	−0.104 (0.024) [0.576]	−0.102 (0.001)	0.017 (0.009)
Weeks worked	−8.71 (0.08)	−5.66 (1.16)	−5.64 (1.16) [0.391]	−8.25 (0.09)	−5.76 (1.15)	−5.76 (1.15) [0.670]	−1.03 (0.05)	1.01 (0.63)
Hours/week	−6.80 (0.07)	−4.08 (0.98)	−4.10 (0.98) [0.489]	−6.39 (0.07)	−3.94 (0.96)	−3.95 (0.96) [0.665]	−0.06 (0.05)	0.85 (0.69)
Labor income	−3984.4 (44.2)	−2099.6 (664.0)	−2096.2 (663.8) [0.830]	−3753.9 (50.7)	−2457.5 (669.7)	−2456.3 (669.7) [0.893]	929.7 (114.9)	1348.7 (1536.0)
ln(*Family income*)	−0.119 (0.005)	−0.124 (0.071)	−0.122 (0.071) [0.270]	−0.103 (0.004)	−0.054 (0.051)	−0.054 (0.051) [0.878]	—	—
ln(*Non-wife income*)	—	—	—	−0.004 (0.005)	0.020 (0.068)	0.020 (0.068) [0.452]	—	—

Notes: The table reports the coefficient on the *More than 2 children* variable in equations (4) and (6) in the text estimated with 1990 Ce data. Other covariates in the models are *Age, Age at first birth,* plus indicators for *Boy 1st, Boy 2nd, Black, Hispanic,* and *Other race.* variable *Boy 2nd* is excluded from equation (6). The p-value for the test of overidentifying restrictions associated with equation (6) is in brackets. Standard errors are reported in parentheses.

The sample was restricted to women under the age of 35 because nearly all children born to women in this age-group are still at home. Relaxing this restriction, a greater fraction of women with two or more children are lost because the oldest child is increasingly unlikely to be at home. We note, however, that the results are not very sensitive to this sample-selection rule. For example, expanding the age-group to include women up to age 45 in the 1980 data, the sample size increases to 552,606 observations. The resulting 2SLS estimate (standard error) of the effect of the *More than 2 children* variable in the *Worked for pay* model is −0.096 (0.021) in the all-women sample.

A final point is that because *Same sex* is an interaction term, the 2SLS estimates were computed using a model that controls for additive effects of s_1 and s_2. This specifica was motivated by a concern with the valid of the instruments and possible omitted va ables bias. It is useful to know whether control for these additive effects is import because if it is, identification turns on our a ity to distinguish additive effects from the teraction term. Moreover, when using *Tw boys* and *Two girls* as separate instruments, must drop one of the additive effects. As turns out, the 2SLS estimates are virtually variant to the inclusion of regressors that co trol for the sex of each child.[13] The coeffici (standard errors) on the *More than 2 child* variable in the *Worked for pay, Weeks wor* and *Hours/week* equations change by no more than 2 percent. There is also little evidence of an association between having male children and labor supply. The only significant sex effects are for s_2 in the 1980 sample, but these are small.[14]

---

[13] The Wald estimates in Table 5 also constitute estimates with no controls for main effects.

E. Heterogeneity in the Impact of Children on Labor Supply

A number of theoretical models describe how the impact of children on labor supply might vary with the wages or schooling of husbands or wives. For example, Angrist and Evans (1996b) outline a version of Gronau's (1977) model of market work and home production that incorporates child quality effects of the sort discussed by Becker and Lewis (1973). This model predicts that higher own wages of either partner magnify the labor-supply consequences of childbearing, although there are no cross-wage effects.[15] Gronau's survey paper (1986 p. 287) refers to a number of empirical studies consistent with this prediction, showing that the labor supply of more educated women is more sensitive to the presence of children than the labor supply of less educated women. Earlier, Gronau (1973 p. S170) reported finding that the effect of a child on his mother's value of time increases with the mother's education. On the other hand, an assumption implicit in most empirical labor-supply models, where the focus is on wage effects and not the consequences of childbearing, is that there are no interactions in the effect of wages and the number of children (see, e.g., Thomas A. Mroz, 1987).

We use the *Same sex* instrument to explore the question of how the labor-market consequences of childbearing varies with the earnings or earnings potential of husbands and wives. Panel A of Table 9 reports OLS and 2SLS estimates of the effect of *More than 2 children* on married women, conditional on the position of their husbands in the husbands' earnings distribution. The first column shows the first-stage relationship between *More than 2 children* and *Same sex,* interacted with dummies that indicate whether husbands' earnings are in the upper third, middle third, or lower third of the earnings distribution. These estimates show that the effect of *Same sex* on fertility is increasing in husbands' earnings. For women with high-wage husbands, however, 2SLS estimates of labor-supply effects are smaller and they are not significantly different from zero. Note that average participation rates do not decline enough with husbands' earnings to account for the decline in the magnitude of the coefficients among women with high-wage husbands. It is also worth noting that the OLS estimates do not decline nearly as much with husbands' earnings as do the 2SLS estimates.

It is not possible to analyze labor-supply effects conditional on women's wages because wages are unobserved for women who do not work. But we can condition on schooling, which is an important predictor of individual earnings potential. This is done in Panel B of Table 9 for married women in the 1980 sample with less than a high-school education (18 percent of the sample), high-school graduates (49 percent of the sample), and more than a high-school education (33 percent of the sample). The reduced forms show a strong association between *Same sex* and fertility in each schooling group, although the effect is about 1 percentage point smaller for mothers in the highest education category. The 2SLS estimates suggest that women with relatively low levels of schooling experience the largest effects of children on labor supply. In contrast, there is no statistically significant association between additional childbearing and labor supply for women with more than a high-school education. As with the estimates that condition on husbands' earnings, the variation in 2SLS estimates by schooling group differs from the variation in OLS estimates,

---

[14] For example, the coefficient (standard error) on *Boy 2nd* in the *Worked for pay* model is −0.0038 (0.0015). Estimates for this variable in the *Weeks worked* and *Hours/week* equations are −0.164 (0.069) and −0.127 (0.059), respectively.

[15] The explanation for this is that in equilibrium, the marginal returns to an hour spent at home are higher for high-wage people than for low-wage people. Although the effects of children are generally ambiguous, in the Angrist and Evans (1996b) model, increasing the number of children increases time spent at home because of returns to scale in parental time spent on child-rearing. Returns to scale are larger when the marginal return to hours (and hence wages) are higher.

TABLE 9—2SLS ESTIMATES OF LABOR-SUPPLY MODELS WITH INTERACTION TERMS USING 1980 CENSUS DATA

Sample/variables	*More than 2 children* First stage	*Worked for pay* Mean of dependent variable	OLS	2SLS	*Weeks/year* Mean of dependent variable	OLS	2SL
A. Results for wives by *husband's earnings:*							
Bottom third of husband's earnings distribution	0.057 (0.003)	0.570	−0.186 (0.003)	−0.122 (0.060)	21.1	−9.23 (0.15)	−7.9 (2.6
Middle third of husband's earnings distribution	0.072 (0.003)	0.569	−0.165 (0.004)	−0.185 (0.047)	20.8	−8.31 (0.15)	−7.11 (2.0
Top third of husband's earnings distribution	0.079 (0.003)	0.448	−0.152 (0.003)	−0.078 (0.042)	15.2	−6.76 (0.15)	−3.11 (1.
B. Results for wives by *wife's education:*							
Wife < high-school graduate	0.071 (0.004)	0.468	−0.150 (0.005)	−0.121 (0.064)	16.1	−7.30 (0.20)	−7.11 (2.8
Wife high-school graduate	0.073 (0.003)	0.524	−0.156 (0.003)	−0.147 (0.038)	19.2	−7.74 (0.13)	−6.4 (1.6
Wife > high-school graduate	0.063 (0.003)	0.567	−0.179 (0.004)	−0.082 (0.054)	20.4	−8.33 (0.15)	−2.9 (2.3
C. Results for wives by *wife's education* for women whose *husband's earnings* are in middle third:							
Wife < high-school graduate	0.079 (0.008)	0.481	−0.138 (0.009)	−0.275 (0.109)	16.7	−7.10 (0.38)	−10.2 (4.
Wife high-school graduate	0.076 (0.004)	0.551	−0.157 (0.003)	−0.189 (0.060)	20.3	−8.33 (0.21)	−7.7 (2.6
Wife > high-school graduate	0.062 (0.006)	0.640	−0.184 (0.006)	−0.125 (0.098)	23.7	−9.07 (0.28)	−3.9 (4.
D. Results for husbands by *wife's education:*							
Wife < high-school graduate	0.071 (0.004)	0.945	−0.014 (0.001)	−0.013 (0.020)	44.5	−1.36 (0.10)	−0.2 (1.3
Wife high-school graduate	0.074 (0.003)	0.981	−0.005 (0.001)	0.005 (0.012)	48.4	−0.53 (0.06)	0.5 (0.5
Wife > high-school graduate	0.063 (0.003)	0.987	−0.002 (0.001)	0.009 (0.016)	49.2	−0.23 (0.08)	0.2 (1.1

Notes: The table reports estimates of the coefficient on *More than 2 children* in equation (4) in the text, modified to al interactions with wives' schooling and husbands' education as indicated. Main effects for each interaction variable (*husban earnings distribution* and *wife's education*) are included in the equation. Other covariates in the models are those listed the notes to Table 7. Data are from the 1980 married women and husband samples. Standard errors are reported in parenthe which show effects that increase in magnitude as schooling increases.

Because mothers' education and husbands' wages are correlated, it is not clear whether a set of estimates that condition on husbands' earnings and a set of estimates that condition on mothers' education are capturing distinct phenomena. We therefore present estimates by mothers' education group in a sample restricted to women whose husbands have earnings in the middle third of the earnings distribution. Again, the 2SLS estimates suggest that the impact of childbearing on labor supply declines as education rises, contradicting a theoretical prediction and the OLS estimates, both of which suggest that the labor supply of more educated women responds more to the presence of children. This finding is even more remarkable when viewed in light of the fact that participation rates increase with mothers' schooling. It should be noted, however, that while the results by education group differ substantially, and the pattern of differences is consistent across outcomes, the estimates for subgroups are not very precise.

The last panel in Table 9 (Panel D) reports estimates for husbands, conditional on wives' education. OLS estimates of labor-supply effects for husbands are small and negative, and they decrease in magnitude as wives' schooling increases. As with the overall estimates for husbands in Table 7, the 2SLS estimates for husbands in Table 9 are small and not significantly different from zero.

Table 10 reports estimates conditional on husbands' earnings and wives' schooling using 1990 data. These results are largely similar to those for 1980, showing 2SLS estimates that decline in magnitude with husbands's earnings and wives' schooling, while the OLS estimates for wives by schooling group are stable or increasing. One interesting difference, however, is that the 1990 results show some small, but statistically significant, *positive* effects of childbearing on the labor supply of the husbands of less educated women. The estimated effect (standard error) on participation rates is 0.031 (0.013) for husbands of women who are high-school graduates and 0.053 (0.023) for husbands of women who did not graduate high school. The effect on weeks worked is also significantly different from zero for the husbands of high-school graduates. These estimates suggest a possible change in husbands' labor-supply response to childbearing between 1980 and 1990, at least for some groups. On the other hand, the 1990 husband effects are still less than half the size of most of the corresponding estimates for women, and they appear even smaller when viewed in light of the greater degree of labor-force attachment among husbands.

III. Comparison with Estimates Using Multiple Births

The most important source of exogenous variation in fertility used in fertility research to date is twinning. Rosenzweig and Wolpin (1980b) used 87 U.S. twin pairs to estimate labor-supply effects, and Rosenzweig and Wolpin (1980a) used 25 twin pairs from India to estimate the effect of family size on school progress. Bronars and Grogger (1994) were the first to study the consequences of multiple births with Census data. They used twins in the 1970 and 1980 PUMS to estimate the effect of additional childbearing on mothers' labor-market status, though most of their estimates are for unwed mothers. Gangadharan and Rosenbloom (1996) also used Census twins to estimate the reduced-form effect of twinning on labor-supply variables, but they fail to scale the reduced-form effects of twinning into effects of childbearing. These studies focused almost exclusively on twinning at first birth. An exception is the Bronars and Grogger study, which also briefly discusses (p. 1149) some estimates using twins at second birth.

A twin second birth is similar to the *Same sex* instrument in that it can be used to measure the consequences of moving from two to three children. We noted in the discussion of Table 4, however, that the use of twins as an instrument may be problematic since twinning probabilities appear to be correlated with some observed characteristics of the mother. On the other hand, if the demographic characteristics associated with twinning are all observed, then these factors can be controlled in 2SLS estimation.

2SLS estimates using *Same sex* and *Twins-2* are compared in Table 11. As in Table 7, the models used to produce these estimates

TABLE 10—2SLS ESTIMATES OF LABOR-SUPPLY MODELS WITH INTERACTION TERMS USING 1990 CENSUS DATA

	More than 2 children	*Worked for pay*			*Weeks/year*	
Sample/variables	First stage	Mean of dependent variable	OLS	2SLS	Mean of dependent variable	OLS
A. Results for wives by *husband's earnings:*						
Bottom third of husband's earnings distribution	0.064 (0.003)	0.668	−0.160 (0.003)	−0.129 (0.045)	26.3	−8.8 (0.15)
Middle third of husband's earnings distribution	0.076 (0.003)	0.728	−0.133 (0.003)	−0.151 (0.039)	29.8	−8.09 (0.15)
Top third of husband's earnings distribution	0.071 (0.003)	0.61	−0.137 (0.003)	−0.029 (0.040)	23.6	−7.27 (0.14)
B. Results for wives by *wife's education:*						
Wife < high-school graduate	0.069 (0.004)	0.531	−0.145 (0.004)	−0.257 (0.061)	19.2	−7.34 (0.20)
Wife high-school graduate	0.078 (0.003)	0.661	−0.140 (0.003)	−0.100 (0.035)	26.3	−8.07 (0.14)
Wife > high-school graduate	0.064 (0.002)	0.718	−0.147 (0.003)	−0.058 (0.038)	29.1	−8.43 (0.13)
C. Results for wives by *wife's education* for women whose *husband's earnings* are in middle third:						
Wife < high-school graduate	0.073 (0.008)	0.579 —	−0.128 (0.008)	−0.279 (0.097)	21.7	−6.92 (0.37)
Wife high-school graduate	0.082 (0.004)	0.707 —	−0.122 (0.005)	−0.204 (0.052)	28.8	−7.62 (0.23)
Wife > high-school graduate	0.071 (0.004)	0.795 —	−0.130 (0.005)	−0.071 (0.060)	33.3	−8.40 (0.28)
D. Results for husbands by *wife's education:*						
Wife < high-school graduate	0.069 (0.004)	0.919 —	−0.033 (0.002)	0.053 (0.023)	42.3	−2.36 (0.11)
Wife high-school graduate	0.076 (0.003)	0.971 —	−0.007 (0.001)	0.031 (0.013)	47.3	−0.70 (0.07)
Wife > high school graduate	0.064 (0.002)	0.982 —	−0.004 (0.001)	−0.014 (0.014)	48.7	−0.41 (0.07)

Notes: The table reports estimates of the coefficient on *More than 2 children* in equation (4) in the text, modified to interactions with wives' schooling and husbands' education as indicated. Main effects for each interaction variable (*husb* *earnings distribution* and *wife's education*) are included in the equation. Other covariates in the models are those list the notes to Table 8. Data are from the 1990 married women and husband samples. Standard errors are reported in parenthe

TABLE 11—COMPARISION OF 2SLS ESTIMATES USING *SAME SEX* AND *TWINS-2* INSTRUMENTS IN 1980 CENSUS DATA

	All women		Married women		Husbands	
Model	(1)	(2)	(1)	(2)	(1)	(2)
Instrument for *More than 2 children*	*Same sex*	*Twins-2*	*Same sex*	*Twins-2*	*Same sex*	*Twins-2*
Dependent variable:						
Worked for pay	−0.125 (0.026)	−0.079 (0.013)	−0.123 (0.028)	−0.087 (0.017)	0.004 (0.009)	−0.001 (0.005)
Weeks worked	−5.82 (1.15)	−3.64 (0.60)	−5.47 (1.23)	−4.21 (0.72)	0.65 (0.61)	−0.35 (0.36)
Hours/week	−4.76 (0.98)	−3.33 (0.51)	−4.91 (1.03)	−3.49 (0.61)	0.57 (0.71)	−0.49 (0.42)
Labor income	−1961.7 (560.5)	−1262.2 (292.8)	−1329.8 (579.1)	−1453.1 (339.8)	−1194.8 (1421.4)	616.8 (836.9)
ln(*Family income*)	−0.021 (0.067)	−0.071 (0.035)	−0.049 (0.057)	−0.025 (0.033)	—	—
ln(*Non-wife income*)	—	—	0.026 (0.068)	0.051 (0.040)	—	—

Notes: The table reports 2SLS estimates of the coefficient on *More than 2 children* in equation (4) in the text using *Same sex* and *Twins-2* as instruments. Other covariates in the models are *Age, Age at first birth,* ages of the first two children, plus indicators for *Boy 1st, Boy 2nd, Black, Hispanic,* and *Other race.* Data are from the 1980 Census. Standard errors are reported in parentheses.

include exogenous covariates to control for mothers' age, race, age at first birth, and the sex of the first two children. Additional covariates included in these models are the ages of the first and second child in quarters. The estimates of female labor-supply effects produced using *Twins-2* are consistently smaller than the corresponding estimates using *Same sex*. Although the contrast between *Same sex* and *Twins-2* coefficient estimates is not large enough to be statistically significant for many of the individual coefficients, the comparison of estimates strongly suggests these two shocks have different effects.

A likely explanation for the smaller *Twins-2* effects is that, conditional on the age of the second child, a third child who is born as a consequence of twinning is necessarily older than a third child who is born for other reasons. This is because third children who are born as twins are exactly the same age as second children, while at least a year and usually longer must go by between the second child's birth and the birth of a non-twin third child. In the 1980 sample, for example, the average age of third children who are twins is 6.4 years while the average age of other third children is five years. Regression-adjusting for the covariates used to construct the estimates in Table 10, the age gap between twins and other third children grows to about 2.5 years. This difference in ages has implications for labor-supply estimates constructed using *Same sex* and twins instruments if the effect of children on labor supply is larger when the children are younger.

We use the following model to check whether differences in the *Same sex* and *Twins-2* 2SLS estimates can be explained by differences in the ages of third children. The equation of interest is

$$(8)\quad y_i = \alpha_0' \mathbf{w}_i + \alpha_1 s_{1i} + \alpha_2 s_{2i} + \alpha_3 a_{1i} + \alpha_4 a_{2i} + \beta_i x_i + \varepsilon_i,$$

TABLE 12—2SLS AND 3SLS ESTIMATES OF TWO-PARAMETER LABOR-SUPPLY MODELS USING 1980 CENSUS DATA

	Worked for pay		*Weeks/year*		*Hours/week*		*Labor income*	
Variable	Not restricted	Restricted	Not restricted	Restricted	Not restricted	Restricted	Not restricted	Restric
A. Instruments: *Same sex* and *Twins-2*								
β_0	−0.191 (0.066)	−0.178 (0.059)	−8.94 (2.91)	−8.24 (2.72)	−6.79 (2.48)	−7.22 (2.38)	−2959 (1423)	[illegible]
β_1	0.015 (0.096)	0.013 (0.009)	0.724 (0.429)	0.616 (0.398)	0.473 (0.366)	0.540 (0.348)	232 (210)	[illegible]
a^*	12.4 (3.69)	13.4 (4.38)	12.3 (3.42)	13.4 (4.38)	14.4 (6.03)	13.4 (4.38)	12.8 (5.62)	[illegible]
B. Instrument: *Same sex* (restricted $a^* = 13.4$)								
β_0	—	−0.184 (0.038)	—	−8.58 (1.69)	—	−7.01 (1.44)	—	[illegible]
C. Instrument: *Twins-2* (restricted $a^* = 13.4$)								
β_0	—	−0.174 (0.030)	—	−8.02 (1.32)	—	−7.35 (1.13)	—	[illegible]

Notes: Panel A of the table reports 2SLS and 3SLS parameter estimates for equation (10) in the text. *Same sex* and *Tw* 2 are both used as instruments. The restricted models in Panel A force the parameter a^* (the age at which labor-sup effects decay to zero) to be the same for all four dependent variables in joint estimation using nonlinear 3SLS. Panel and C report 2SLS estimates of equation (11) using the *Same sex* and *Twins-2* instruments separately. Other covari in the models are listed in the notes to Table 11. The data are from the 1980 Census all-women sample. Standard er are reported in parentheses.

where a_{1i} and a_{2i} are the ages of the first two children. The coefficient β_i is now an individually varying causal effect that depends on the age of the third child. In particular, we assume

$$(9) \qquad \beta_i = \beta_0 + \beta_1 a_{3i},$$

where a_{3i} is equal to the age of the mothers' third child for women who have a third child and is equal to zero otherwise. Combining (8) and (9) generates the estimating equation,

$$(10) \quad y_i = \alpha_0' \mathbf{w}_i + \alpha_1 s_{1i} + \alpha_2 s_{2i} + \alpha_3 a_{1i}$$
$$+ \alpha_4 a_{2i} + \beta_0 x_i + \beta_1 (a_{3i} x_i) + \varepsilon_i.$$

Assuming that differences in a_{3i} are the only reason why the *Same sex* and *Twins-2* instruments generate different estimates, we can use both instruments to estimate the coefficients on the two endogenous regressors in (10), and $a_{3i} x_i$.

2SLS estimates of β_0 and β_1 are reported Table 12 for the full 1980 sample, where was measured to the nearest quarter for purposes of estimation. All of the estimates β_0 are negative and all of the estimates of are positive, suggesting that the negative pact of childbearing declines as the third ages. The table also reports estimates of value of a_{3i} at which $\beta_i = 0$; this is $a^* \equiv$ β_1. Estimates of a^* are 12.4 years for eff on *Worked for pay,* 12.3 years for effects *Weeks worked,* 14.4 years for effects *Hours/week,* and 12.8 years for effects on *bor income.*[16] We also estimated a^* under restriction that $\beta_1 a^* \equiv -\beta_0$ with the same value of a^* across all four dependent variables. The estimation method for this model is three-stage least squares (3SLS). The restricted estimate is 13.4 with standard error of 1.4. Imposing this restriction leads to slightly smaller standard errors for the estimates of β_0 and β_1. The test statistic for this restriction, distributed as $\chi^2(3)$ under the null hypothesis that the restrictions are satisfied, takes on the value 0.79, which has a p-value of about 0.86.

---

[16] The linear model for β_i is obviously an appro tion since it implies that the effects of childbear

To further illustrate how this model reconciles the *Same sex* and *Twins-2* estimates, note that if $\beta_1 = -\beta_0/a^*$, we have

$$(11)\quad y_i = \alpha_0' \mathbf{w}_i + \alpha_1 s_{1i} + \alpha_2 s_{2i} + \alpha_3 a_{1i} + \alpha_4 a_{2i} + \beta_0(1 - a_{3i} / a^*)x_i + \varepsilon_i .$$

Substituting the pooled estimate of a^* into (11), we can use the *Same sex* and *Twins-2* instruments to construct separate estimates of β_0 in (11) by treating $(1 - a_{3i}/a^*)x_i$ as the single endogenous regressor. The results when $a^* = 13.4$, reported in Panels B and C of Table 12, show that the *Same sex* and *Twins-2* instruments generate very similar estimates of β_0 for all dependent variables. This suggests that the model of the effect of childbearing embodied in (9), combined with the restriction that β_t decays to zero at age 13.4, does a good job of reconciling the *Same sex* and *Twins-2* estimates.[17]

Differences in the age of the third child constitute one of many possible explanations for the contrast between the *Same sex* and *Twins-2* estimates. For example, there may be economies of scale in parenting two children of the same age. On the other hand, closely spaced young children may require more attention than an older child and a younger child. It is worth noting, however, that the model outlined in this section also serves to explain why the twins estimates reported by Bronars and Grogger (1994) for married mothers are smaller than the *Same sex* estimates reported here. When Bronars and Grogger use twins to estimate the effects of childbearing conditional on the age of the first child (and hence on the age of the twin), they find effects on the labor-force participation rates of mothers of children aged 0–3 remarkably similar to the *Same sex* estimates, with no effects for the mothers of children aged 10–13 (see Table 4 in Bronars and Grogger, 1994). Because second-born twins are younger, on average, than firstborn twins, age differences could also explain why the effects of twins at second birth briefly mentioned by Bronars and Grogger in the text of their article (p. 1149) are larger than their estimates of the effects of twins at first birth.[18]

IV. Implications for the Increase in Female Labor Supply

At the turn of the century, less than 20 percent of all workers were women. Today, women make up almost half the workforce (Goldin, 1990). A number of researchers have attempted to decompose the rise in the female labor-force participation rates into components attributable to demand and supply shifts. For example, Mincer (1962) concluded that 90 percent of the rise in postwar labor-force participation of married women can be attributed to an increase in demand. James P. Smith and Ward (1984) also found that demand characteristics can explain a majority of the increase in total hours worked by all women in between 1850 and 1980. In contrast, Goldin (1990) argues that shifts in supply explain about half of the change in female

---

labor supply become positive once the third child is older than a^*. This approximation seems harmless since only about 2 percent of third children are older than 13 in our data (because the oldest mother is aged 35).

[17] If we set $a^* = -\beta_0/\beta_1$ using the coefficient estimates from each equation, then the *Same sex* and *Twins-2* estimates of (11) for any equation will necessarily be identical. The point of estimation with a^* fixed at 13.4 is to show how one extra free parameter reconciles *all four* of the *Same sex* and *Twins-2* estimates.

[18] The fact that the Bronars and Grogger estimates are less precise than the twins estimates reported here is likely due to their having drawn a 1-percent sample of singleton births for the comparison sample. Note also that Bronars and Grogger's estimates are for the reduced-form impact of twinning; for purposes of comparison, these estimates should be scaled up by dividing by the twins first-stage effect (about 0.68 in their data).

labor-force participation between 1960 and 1980.

Declining fertility represents a potentially important supply shifter that might account for some of the increase in female labor-force attachment. How much of the trend in labor-force attachment in the population we have studied can be accounted for by reduced childbearing beyond the second child? Table 1 showed that the probability of having more than two children for women aged 21–35 with at least two children fell by 16.5 percentage points between 1970 and 1990, a drop of about 30 percent. At the same time, labor-force participation rates rose by 21.8 percentage points, a 49-percent increase. Similar statistics for other groups reported in Table 1 show that our sample is not unusual in experiencing these trends.

Using the *Same sex* 2SLS estimate of the impact of *More than 2 children* on *Worked for pay* from Table 7 (−0.119), declining fertility can account for an employment increase equal to 0.165 × 0.12, which is about 2 percentage points. This calculation suggests that even though childbearing clearly affects labor supply, the increase in female labor-force participation has been so large that declining fertility can explain only a small fraction of the overall change.

V. Conclusions

Economic models of household behavior generate a rich variety of predictions and theoretical relationships, few of which have been confronted with credible empirical evidence. The evidence reported here is unique in that it derives from plausibly exogenous sources of variation in family size. However, the empirical results probably raise as many questions as they answer.

2SLS and IV estimates that exploit the fertility consequences of sibling sex composition and twinning both confirm the OLS estimates showing that children lead to a reduction in female labor supply, although the OLS estimates appear to exaggerate the causal effect of children. This is probably not too surprising, at least not to the mothers of small children. What is surprising is that the effects of children on labor supply appear to be much smaller and possibly even ab among college-educated women and wo whose husbands have high wages. This re contradicts the predictions of some theo of household time allocation as well as t OLS results, which suggest that the lab supply consequences of childbearing larger for more educated women. Our es mates consistently show that the labor-mark consequences of childbearing are more lik to be severe for poor and less educat women.

Equally important is the finding that hu bands change their labor-market behavi very little in response to a change in fami size. Even the husbands of well-educated a relatively well-paid women do not cha their work habits in response to the birth o child. Thus, families absorb the cost of car for a third child either through a reduction wife's earnings or by purchasing child-ca services from nonfamily providers. If ad tional childbearing does lead husbands to p additional time into home production of ch care, this is done at the expense of the hu band's leisure time and not through a red tion in his work effort. We also find li evidence of an increase in husbands' earni that would offset the decline in wives' ea ings. In spite of the increase in wom wages and labor-force participation rates d ing the period studied here, the labor-mark behavior of most married men appears t have remained largely insensitive to the nu ber of children.

REFERENCES

Angrist, Joshua D. "Grouped-Data Estima and Testing in Simple Labor Supply M els." *Journal of Econometrics*, Febru March 1991, *47*(2–3), pp. 243–66.

Angrist, Joshua D. and Evans, Willia "Schooling and Labor-Market Co quences of the 1970 State Abortion R forms." National Bureau of Econo Research (Cambridge, MA) Working P No. 5406, January 1996a.

______. "Children and Their Parents' L Supply: Evidence from Exogenous V tion in Family Size." National Bure Economic Research (Cambridge, Working Paper No. 5778, September 1996b.

Angrist, Joshua D. and Lavy, Victor. ''The Effect of Teen Childbearing and Single Parenthood on Childhood Disabilities and Progress in School.'' National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5807, October 1996.

Ashenfelter, Orley and Heckman, James J. ''The Estimation of Income and Substitution Effects in a Model of Family Labor Supply.'' *Econometrica*, January 1974, *42*(1), pp. 73–85.

Becker, Gary S. ''Human Capital, Effort, and the Sexual Division of Labor.'' *Journal of Labor Economics*, January 1985, Pt. II, *3*(1), pp. S33–58.

Becker, Gary S.; Landes, Elisabeth M. and Michael, Robert T. ''An Economic Analysis of Marital Instability.'' *Journal of Political Economy*, December 1977, *85*(6), pp. 1141–87.

Becker, Gary S. and Lewis, Gregg H. ''On the Interaction Between the Quantity and Quality of Children.'' *Journal of Political Economy*, March–April 1973, Pt. II, *81*(2), pp. S279–88.

Becker, Gary S. and Tomes, Nigel. ''Child Endowments and the Quantity and Quality of Children.'' *Journal of Political Economy*, August 1976, Pt. 2, *84*(4), pp. S143–62.

Ben-Porath, Yoram and Welch, Finis. ''Do Sex Preferences Really Matter?'' *Quarterly Journal of Economics*, May 1976, *90*(2), pp. 285–307.

______. ''On Sex Preferences and Family Size,'' in Julian L. Simon and Julie DeVanzo, eds., *Research in population economics*. Greenwich, CT: JAI Press, 1980, pp. 387–99.

Blau, Francine and Grossberg, Adam J. ''Maternal Labor Supply and Children's Cognitive Development.'' *Review of Economics and Statistics*, August 1992, *74*(3), pp. 474–81.

Bronars, Stephen G. and Grogger, Jeff. ''The Economic Consequences of Unwed Motherhood: Using Twins as a Natural Experiment.'' *American Economic Review*, December 1994, *84*(5), pp. 1141–56.

Browning, Martin. ''Children and Household Economic Behavior.'' *Journal of Economic Literature*, September 1992, *30*(3), pp. 1434–75.

Butcher, Kristin F. and Case, Anne. ''The Effect of Sibling Sex Composition on Women's Education and Earnings.'' *Quarterly Journal of Economics*, August 1994, *109*(3), pp. 531–63.

Butz, William P. and Ward, Michael. P. ''The Emergence of Countercyclical U.S. Fertility.'' *American Economic Review*, June 1979, *69*(3), pp. 318–28.

Coleman, Mary T. and Pencavel, John. ''Trends in Market Work Behavior of Women Since 1940.'' *Industrial and Labor Relations Review*, July 1993, *46*(4), pp. 653–76.

Gangadharan, Jaisri and Rosenbloom, Joshua L. ''The Effects of Child-bearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment.'' National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5647, July 1996.

Goldin, Claudia. *Understanding the gender gap*. New York: Oxford University Press, 1990.

______. ''Career and Family: College Women Look to the Past.'' National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5188, July 1995.

Gronau, Reuben. ''The Effect of Children on the Housewife's Value of Time.'' *Journal of Political Economy*, March–April 1973, Pt. II, *81*(2), pp. S168–99.

______. ''Leisure, Home Production, and Work—The Theory of the Allocation of Time Revisited.'' *Journal of Political Economy*, December 1977, *85*(6), pp. 1099–124.

______. ''Home Production—A Survey,'' in Orley Ashenfelter and Richard Layard, eds., *Handbook of labor economics*, Vol. 1. Amsterdam: North-Holland, 1986, pp. 273–304.

______. ''Sex-Related Wage Differentials and Women's Interrupted Careers—The Chicken or the Egg.'' *Journal of Labor Economics*, July 1988, *6*(3), pp. 277–301.

Gruber, Jonathan and Cullen, Julie Berry. ''Spousal Labor Supply as Insurance: Does Unemployment Insurance Crowd Out the Added Worker Effect?'' National Bureau of

Economic Research (Cambridge, MA) Working Paper No. 5608, June 1996.

Heckman, James J. and MaCurdy, Thomas E. "A Life-Cycle Model of Female Labor Supply." *Review of Economic Studies*, January 1980, *47*(1), pp. 47–74.

Heckman, James J. and Walker, James R. "The Relationship Between Wages and Income and the Timing and Spacing of Births: Evidence from Swedish Longitudinal Data." *Econometrica*, November 1990, *58*(6), pp. 1411–441.

Iacovou, Maria. "Fertility and Female Labor Force Participation." Unpublished manuscript, University College, London, July 1996.

Imbens, Guido W. and Angrist, Joshua D. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, March 1994, *62*(2), pp. 467–76.

Killingsworth, Mark R. and Heckman, James J. "Female Labor Supply: A Survey," in Orley Ashenfelter and Richard Layard, eds., *Handbook of labor economics*, Vol. 1. Amsterdam: North-Holland, 1986, pp. 103–204.

Korenman, Sanders and Neumark, David. "Marriage, Motherhood, and Wages." *Journal of Human Resources*, Spring 1992, *27*(2), pp. 233–55.

Mincer, Jacob. "Labor Force Participation of Married Women: A Study of Labor Supply," in H. Gregg Lewis, ed., *Aspects of labor economics*. Princeton, NJ: Princeton University Press, 1962, pp. 63–105.

______. "Market Prices, Opportunity Costs, and Income Effects," in Carl F. Christ, ed., *Measurement in economics*. Stanford, CA: Stanford University Press, 1963, pp. 67–82.

Montgomery, Mark R. and Trussell, James. "Models of Marital Status and Childbearing," in Orley Ashenfelter and Richard Layard, eds., *Handbook of labor economics*, Vol. 1. Amsterdam: North-Holland, 1986, pp. 205–71.

Morgan, Philip S.; Lye, Diane N. and Condran, Gretchen A. "Sons, Daughters, and the Risk of Marital Disruption." *American Journal of Sociology*, July 1988, *94*(1), pp. 110–29.

Mroz, Thomas A. "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical sumptions." *Econometrica*, July 1 *55*(4), pp. 765–99.

Myrianthopoulos, Ntinos C. "An Epidemio Survey of Twins in a Large Prospecti Studied Population." *American Journa Human Genetics*, November 1970, 22 pp. 611–29.

Nakamura, Alice and Nakamura, Masao. Econometrics of Female Labor Supply Children." *Econometric Reviews*, 19 *11*(1), pp. 1–71.

National Center for Health Statistics. *Vital sta tics of the United States: Natality*. Hya ville, MD: National Center for Hea Statistics, various years.

Newey, Whitney K. and West, Kenneth D. pothesis Testing with Efficient Metho Moments Estimation." *International E nomic Review*, October 1987, *28*(3), 777–87.

Pencavel, John. "Labor Supply of Men: A S vey," in Orley Ashenfelter and Rich Layard, eds., *Handbook of labor econ ics*. Vol. 1. Amsterdam: North-Holl 1986, pp. 3–101.

Rosenzweig, Mark R. and Wolpin, Kenne "Testing the Quantity-Quality Fert Model: The Use of Twins as a Natural periment." *Econometrica*, January 19 *48*(1), pp. 227–40.

______. "Life-Cycle Labor Supply and tility: Causal Inferences From Hous Models." *Journal of Political Econ* April 1980b, *88*(2), pp. 328–48.

Schultz, Paul T. *Economics of populat* Reading, MA: Addison-Wesley, 1981.

______. "Testing the Neoclassical Mod Family Labor Supply and Fertility." *nal of Human Resources*, Fall 1990, 2 pp. 599–634.

Smith, James P. and Ward, Michael P. *Wo wages and work in the twentieth cen* Santa Monica, CA: Rand Corpor 1984.

Stafford, Frank P. "Women's Work, S Competition, and Children's School P mance." *American Economic Review*, cember 1987, *77*(5), pp. 972–80.

Thomas, Duncan. "Like Father, Like Like Mother, Like Daughter: Parent sources and Child Height." *Jour*

Human Resources, Fall 1994, *29*(4), pp. 950–89.

U.S. Department of Commerce, Bureau of the Census. *Census of population and housing, 1970 [United States]: Public use microdata samples 1/100, 5-percent file* [computer file]. Washington, DC: U.S. Department of Commerce, Bureau of the Census [producer], 1972.

———. *Census of population and housing, 1980 [United States]: Public use microdata sample, 5-percent sample* [computer file]. Washington, DC: U.S. Department of Commerce, Bureau of the Census [producer], 1983.

———. *Census of population and housing, 1990 [United States]: Public use microdata sample, 5-percent sample* [computer file]. Washington, DC: U.S. Department of Commerce, Bureau of the Census [producer], 1995.

Waterhouse, John A. H. "Twinning in Twin Pedigrees." *British Journal of Social Medicine*, October 1950, *5*, p. 197.

Westoff, Charles F.; Potter, Robert G. and Sagi, Phillip C. *The third child: A study in the prediction of fertility*. Princeton, NJ: Princeton University Press, 1963.

Williamson, Nancy E. *Sons or daughters: A cross-cultural survey of parental preferences*. Beverly Hills, CA: Sage, 1976.

Willis, Robert J. "What Have We Learned from the Economics of the Family?" *American Economic Review*, May 1987 (*Papers and Proceedings*), *77*(2), pp. 68–81.

Does European Unemployment Prop Up American Wages? National Labor Markets and Global Trade

By DONALD R. DAVIS*

I consider trade between a flexible-wage America and a rigid-wage Europe. In a benchmark case, a move from autarky to free trade doubles European unemployment. American wages rise to the European level. Entry of the unskilled "South" to world markets raises European unemployment. Europe's commitment to the high wage wholly insulates America from the shock. Immigration to America raises American income, but lowers European income dollar for dollar, while European unemployment rises. Absent South–North migration of the unskilled from 1970–1990, Europe could have maintained the same wage with from one-eighth to one-fourth less unemployment. (*JEL* F11, J31, E24)

I. A Global Approach

In recent years, factor market developments in the United States and United Kingdom have contrasted sharply with those in continental Europe. In the United States and United Kingdom, the relative wage of the unskilled has declined significantly. In the span of a decade (1979–1989) the wage of a U.S. worker in the 90th percentile relative to that of one in the 10th percentile rose by over 20 percent (Richard B. Freeman and Lawrence F. Katz, 1995). This rising wage inequality was much less evident in continental Europe. However, unemployment has risen sharply in Europe. The early postwar decades are now thought of as a "Golden Age" for Europe, with unemployment rates of 2 to 3 percent. Beginning in the 1970's these rates have climbed dramatically, reaching double digits in many of these countries (Center for Economic Policy Research [CEPR], 1995).

An extensive empirical literature has sidered the provenance of these factor m developments. The studies are of two pri pal types (Freeman and Katz, 1995). The dividual country studies provide a account of local developments in inst tions, factor supplies, and demand condit The comparative (cross-country) studies stract from local idiosyncracies to search common themes. However, even the com ative studies suffer from an important d back: they remain a collection of indiv stories. They do not pretend to provide a c mon framework—and the consistency enforces—to provide a unified account

This suggests the value of a third app which may be termed "global." It is, in first instance, a general equilibrium s Consistency is enforced by the fact that is a simultaneous determination of equilib in all of the factor markets. Naturally, tr bility limits the degree of local institution tail that may be considered. Considering important differences in factor market in tions will nonetheless be an important f of such an approach.

However, a global approach is more just general equilibrium. It aims at a explanation of very divergent exper There are two reasons for seeking such fied account. First, many of the shocks the industrial "North" are common

---

[*] Department of Economics, Harvard University, Cambridge, MA 02138, and National Bureau of Economic Research. I would like to acknowledge, without implication, the helpful comments of Richard Brecher, Carolyn Evans, John Leahy, Howard Shatz, David Weinstein, and Jeff Williamson. I am also grateful for the support for this project from the Harvard Institute for International Development (HIID).

example, the entry to the integrated world economy of important new trading partners from the newly industrializing "South."[1] If we are to believe such accounts, we need to see how these shocks interact with a variety of local institutions. A more subtle—but by this more important—reason for considering a global approach is that the consequences even of purely local institutions and shocks often depend crucially on the links to the global market.

The paper joins two elements that have figured prominently in these discussions, but whose implications have not previously been worked out in a consistent manner. The first is the stylized characterization of America as a flexible-wage economy, and Europe as an economy in which a variety of institutions—unions, explicit minimum wages, etc.—make wages more rigid.[2] The second is the Heckscher-Ohlin model, widely acknowledged to be the appropriate framework for examining movements in relative wages (Paul R. Krugman, 1996).

Thus, we will build a model of world trade between a flexible-wage America and a minimum-wage Europe. In doing so, we build on the classic work of Richard A. Brecher (1974a, b). The America versus Europe dichotomy has been featured in a variety of recent work.[3] The distinctive feature of the present work is the explicit focus on factor markets in *both* countries, and particularly on their interaction owing to their link via trade in goods.

This paper confirms that national factor market institutions matter. They profoundly affect global patterns of output, employment, and wages. Surprisingly, though, in the present context they have no power in accounting for cross-country differences in the evolution of relative wages. Often forgotten in this context is that even countries with distinctive factor market institutions are linked by world commodity prices and producers' zero profit conditions. Hence an account of the observed wage trends must move beyond an appeal to local institutional differences.

This paper is comprised of seven parts. Section II examines the consequences of differing labor-market institutions for countries linked by world commodity trade in homogeneous goods. Section III examines the contrasting impacts on the stylized America and Europe posed by the appearance of newly industrializing countries. Section IV examines the effects of factor accumulation in the two countries on wages and unemployment. Section V considers the implications of South–North migration for unemployment in Europe. Section VI returns to the issue of divergent wage experience between America and Europe. Section VII concludes.

II. National Labor Markets and Global Trade

A. *Unemployment in the Global Equilibrium*

The aim of this paper is to develop a model of trade between two countries, one of which has flexible wages, while the other imposes a binding minimum wage on unskilled labor. It is convenient to develop this in three stages. The first considers a conventional two-sector general equilibrium model of a closed economy and establishes some relations that are key to later analysis. The second introduces a binding minimum wage within the closed economy. Finally, I show the isomorphism between this economy and the two-country trading world of interest.[4]

---

[1] Another source of common shocks—those due to global technological change—is analyzed in Davis (1998).

[2] There are alternative approaches to introducing unemployment into general equilibrium trade models. For example, Steven J. Matusz (1996) develops an elegant model merging elements of efficiency-wage unemployment with monopolistic competition to show how the increased market size due to trade can raise real wages, so reduce equilibrium unemployment in both countries. This provides insight for questions such as European integration. The simpler model that is developed here, though, captures the essential elements of relative demand shifts that have been seen as the heart of recent developments in both Europe and America. See, e.g., Kevin M. Murphy (1995).

[3] This includes Giuseppe Bertola and Andrea Ichino (1995), Freeman and Katz (1995), Krugman (1994, 1995), and Adrian Wood (1994). Bertola and Ichino develop an insightful approach that identifies rising volatility of idiosyncratic shocks as a potential account of the growing wage dispersion in the United States and rising unemployment in Europe. Their focus on homogeneous labor suggests a story of growing within-group dispersion. The remaining papers are closer to the approach of the present paper, focusing on heterogeneous labor. Here I provide a consistent general equilibrium model integrating the flexible/rigid dichotomy in the context of heterogeneous labor.

Consider a closed flexible-wage economy. There are two factors of production: skill and labor (also termed the unskilled). They are available in fixed supply given respectively by H^W and L^W (shortly these will be the world endowments of a two-country world). These are used to produce two goods under constant returns to scale. Assume that at any common factor prices, good X is skill intensive relative to Y. Assume that preferences are homothetic, and that both goods are necessary in consumption. Let w be the return to labor, r the return to skill, and P the relative price of X in terms of Y. Then the competitive cost conditions insure that for each sector price equals unit cost:

$$(1) \qquad c_X(w, r) = P \qquad c_Y(w, r) = 1.$$

Assume that marginal products are always strictly positive. Then flexible wages insure full employment. Factor market clearing requires that employment in the two sectors equal the world endowment:

$$(2) \qquad H_X + H_Y = H^W \qquad L_X + L_Y = L^W.$$

By Walras' law, goods market clearing is insured by equality of demand and supply of good X.

Three relations are key to our analysis. The first is a relation between the price P in a closed economy and the endowment ratio, $h^W \equiv H^W/L^W$, of that economy. The Heckscher-Ohlin theorem insures that a rise in skill abundance reduces the relative price of the skill-intensive good. This relation expressed as:

$$(3) \qquad P = \lambda(h), \text{ where } \lambda'(h) < 0.$$

The second is the Stolper-Samuelson tion between goods prices and factor pr Given the assumption that X is alway skill-intensive sector, this defines a mono relation, of which I focus on only a part, Stolper-Samuelson:

$$(4) \qquad w = \psi(P), \text{ where } \psi'(P) < 0.$$

Thus given the endowment ratio of the c economy, h^W, one can determine directly equilibrium goods price in the flexible- case, $P^F = \lambda(h^W)$. Likewise, one can d the resulting flexible wage, $w^F = \psi(P$ $\psi(\lambda(h^W))$, as well as the associated s wage. These suffice to establish the basic acteristics of the world equilibrium.

I now introduce within the closed econ a binding minimum wage for labor at $w^* > w^F$. This will be consistent with an librium featuring diversified production if only if the relative goods price is P $\psi^{-1}(w^*) < P^F$. However, these will be equilibrium goods and factor prices if and if *employed* factors are in the ratio h $\lambda^{-1}(\psi^{-1}(w^*)) > h^W$. The flexible rental skill insures that it will always be fully ployed. However with a binding mini wage, this need not be true for labor. Th manner in which relative employed facto to h^* is for the denominator to fall vi employed labor. With w^* given, h^* is mined, and H^W and L^W are the fixed endowments. Let N be the level of actually employed. Then simple al shows that unemployment in this eco $U = L^W - N$, is given by the third key r (Brecher):

$$(5) \qquad U^* = L^W - H^W/h^*$$
$$\equiv \beta(h^*; H^W, L^W),$$
$$\text{where } \beta'(h^*) > 0.$$

These three key relations characteri minimum-wage economy can be cons

---

[4] Two additional papers have independently considered two-country models with unemployment. N. Boccard et al. (1996) develop a political economy model focused on redistribution and "social dumping." Paul Oslington (1996) considers trade patterns and "false" comparative advantage in a factor price equalization (FPE) setting, suggesting that transport costs may resolve indeterminacies in the cross-country distribution of unemployment.

Figure 1. The Heckscher-Ohlin relation, $P = \lambda(h)$, appears in quadrant one. The Stolper-Samuelson relation, $w = \psi(P)$, appears in quadrant two. Finally, the link between endowments, the employment ratio, and the level of unemployment is depicted in quadrant four. In the flexible-wage case, the employment ratio equals the endowment ratio h^W. This determines wages, prices, and unemployment as w^F, P^F, and zero, respectively. In the minimum-wage case, the path of determination is different. The given minimum wage $w^* > w^F$ directly determines the goods price, the employment ratio, and the unemployment level as P^*, h^*, and U^*, respectively.

The basic link between the minimum wage and the appearance of unemployment in this economy is very simple. For competitive firms to pay the high wage w^*, this must be supported by an appropriate goods price P^*. When the minimum wage binds, the appropriate goods price will be attained only if the relative scarcity of the labor-intensive good rises relative to the flexible-wage equilibrium. And this will occur only if a sufficient amount of labor is unemployed.

I now establish an isomorphism between the closed economy with a minimum wage and a trading world in which one economy imposes a minimum wage while the other maintains a flexible wage. This relies on a concept introduced by Avinash K. Dixit and Victor F. Norman (1980) known as the integrated equilibrium. This establishes conditions under which trade in goods alone suffices to establish the same world equilibrium as occurs in the closed economy with both goods and factor mobility.

Consider, then, a world of two countries with free trade and zero transport costs. One country—America—has flexible wages. The other country—Europe—has imposed a minimum wage at the level w^*. The technologies and preferences in the two countries are identical to those in the closed economy. The technologies are constant returns to scale, while the preferences are identical and homothetic. Let a bar over a variable represent the level of that variable in the integrated equilibrium. Let i index goods and j index countries. The set of divisions of world endowments among the two countries consistent with replicating the integrated equilibrium can then be described. This is called the FPE set, and is described as:

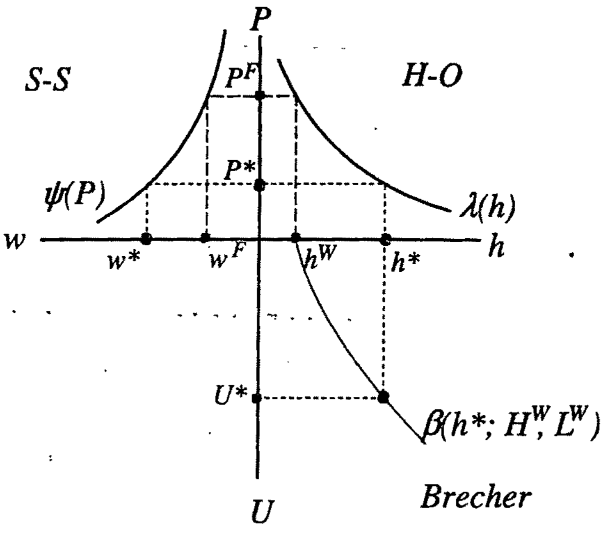


FIGURE 1. UNEMPLOYMENT IN GENERAL EQUILIBRIUM

$$\mathrm{FPE} = \left\{ \begin{array}{l} [(H^A, L^A), (H^E, L^E)] \mid \exists \lambda_{ij} \geq 0 \\ \quad \text{such that} \sum_j \lambda_{ij} = 1, \\ (H^E, L^E) = \sum_i \lambda_{i,E}(\bar{H}(i), \bar{L}(i)) \\ \qquad + (0, U^*) \\ (H^A, L^A) = \sum_i \lambda_{i,A}(\bar{H}(i), \bar{L}(i)) \\ \qquad i = X, Y \quad j = A, E \end{array} \right\}$$

These conditions are very intuitive. If the integrated equilibrium is to be replicated, then world unemployment must be at the same level as in the integrated economy. But unemployment cannot arise in the flexible-wage America. Hence Europe must endure the entire integrated equilibrium level of unemployment, U^*, to maintain the wage w^*. Beyond this, we need only satisfy the conventional restrictions in terms of *employed* factors. These require that both countries use the integrated equilibrium techniques (with the skill intensities $h_X^* = H_X^*/L_X^*$ and $h_Y^* = H_Y^*/L_Y^*$), that the integrated equilibrium output in both sectors be divided among the countries, and that this exactly exhaust employed factors in the two countries (which exist in the ratio h^*).[5] The FPE set is depicted as the bold parallelogram in Figure 2. The level of world (hence European) unemployment is indicated by the line segment O^EU.

This allows me to state a key result: Under the conditions noted above, international trade equalizes factor prices between the flexible-wage and the minimum-wage economies. The proof is simply that under free and costless trade, competitive producers in the two countries face the same goods prices, have the same technologies, and are (at least weakly) diversified. The equality of factor prices then follows directly from the common competitive cost conditions.

This is an important result. Even in the face of sharply divergent factor market institutions, free commodity trade in a global market fully equalized factor prices. As will be considered more fully below, this will sharply restrict the nature of shocks that can be appealed to in explaining divergent wage trends between Europe and America.[6]

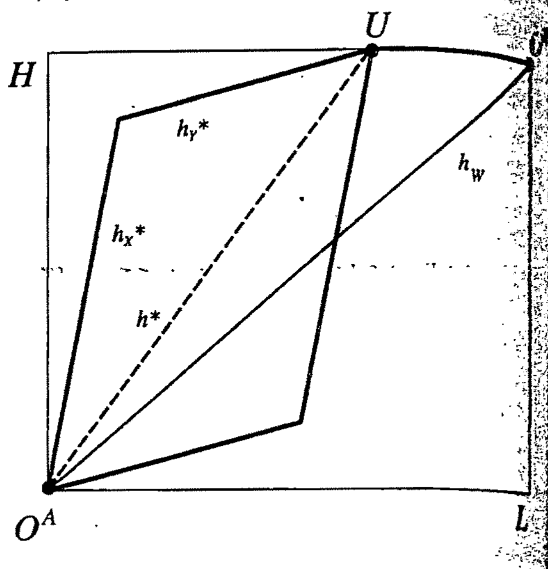


FIGURE 2. UNEMPLOYMENT IN THE INTEGRATED WORLD ECONOMY

B. *Trade and Unemployment*

I have shown how to construct a mode
trade between a flexible-wage economy
one with a minimum wage. But I have not
examined what impact trade has on th
economies relative to autarky. It is conve
to start with a highly stylized example that
tablishes a benchmark. Further insight into
workings of this model will be provided by
many comparative statics in later sections.

Take as a benchmark a world in which
rope and America are alike in every resp
endowments, technologies, and prefere
The one exception is that America has a
ible wage, while Europe has a minimum
at level w^*, assumed to bind in autarky.
before, technologies are constant returns
scale. For simplicity, assume that prefer
of the representative consumer are homot

---

[5] Since preferences are assumed to be identical and homothetic, the level of spending by the unemployed does not matter for the pattern of world spending. Implicitly, though, it is assumed that any spending by the unemployed is financed via lump-sum transfers.

[6] For analytic simplicity, I developed this within a conventional Heckscher-Ohlin framework. One feature of this model—that American and European goods are perfectly homogeneous—may appear crucial to the results on FPE and consequent international spillovers to European unemployment. However, such an impression would be incorrect, at least in the longer run with which this model is concerned. It would be straightforward, although not as clean, to consider a model in which the homogeneous goods X and Y were replaced by two monopolistically competitive industries, $\{X\}$ and $\{Y\}$, each with many varieties (see Elhanan Helpman and Krugman, 1985). In this case, all goods, wherever produced, are imperfect substitutes for all others. Yet under the same institutional assumptions, one could still apply precisely the same integrated equilibrium approach to again demonstrate factor price equalization with unemployment in Europe alone. What is crucial for the strong result is that European and American labor of the various categories are effectively homogeneous. In such a case, a disturbance may in the short run introduce a divergence from unity in the relative price of European and American varieties. This
give rise to cross-country factor price inequalities.
ever in the long run, this will induce entry, exit, and
adjustments that return us to the integrated equili
and restore full factor price equalization. Thus, to
from this paper's results in the long run, it does not
to observe that American and European goods are
fect substitutes. One must believe that on a deep
American and European labor of the various class
different factors.

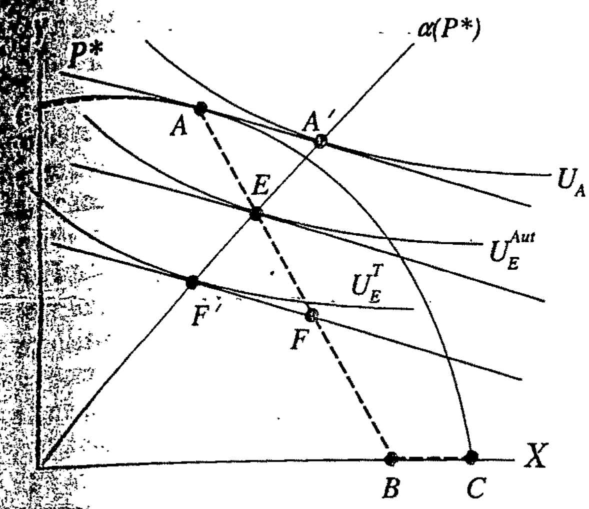


FIGURE 3. BENCHMARK: FREE TRADE DOUBLES EUROPEAN UNEMPLOYMENT

Figure 3 depicts the salient aspects of the equilibrium.[7] The technological production frontier, CC', is common to America and Europe. In Europe, to support the wage w^*, the equilibrium price must be P^*. The consumption ratio of Y to X at P^* is given by $\alpha(P^*)$. Thus under full employment, production at prices P^* in Europe would be at A, and demand at A'. This implies an incipient excess demand for X, tending to raise P and lower w below w^*. The incipient fall in the wage is instead staunched via a decline in unskilled employment in Europe. With goods prices fixed at P^*, this is just the Rybczynski theorem in reverse, shifting output along a minimum wage-constrained production possibility frontier (PPF) indicated by AB. Autarky equilibrium in Europe is at E, at which the constrained supply exactly matches the demand at prices P^*. The shift of production in Europe from A to E reflects the contraction of employment just necessary to eliminate the excess demand indicated by AA'. By contrast, in America, the autarky equilibrium features a price $P > P^*$ and so an unskilled wage $w < w^*$ (not depicted). There is no unemployment in America.

Now consider the equilibrium when America and Europe trade freely. In order to support the minimum wage w^*, Europe still needs to maintain the price P^* in equilibrium. At prices P^*, the American net trade offer is indicated by AA'. Thus world equilibrium at these prices requires that Europe be willing to make this trade at prices P^*. This requires that European production be at point F, and absorption at F'. As demand is homothetic, and production linear at all points along AB, the loss of employment in moving from E to F matches the preexisting unemployment indexed by AE.[8]

This implies a striking contrast between Europe and America in the labor market. Europe's commitment to a high wage gave rise to unemployment in autarky. In this stylized example the opening of Europe to trade with America *doubles* European unemployment, as it is forced to absorb the full integrated equilibrium level of unemployment to sustain w^* for both. In America, the absence of unemployment in autarky came at the cost to workers of a lower real wage. However once trade commenced, American workers came to share the high European wage even as they suffered none of the unemployment that sustains that wage. The fact that they suffer no unemployment is a consequence of the flexible wage in America. The fact that they nonetheless share the European high wage under trade follows from the fact that trade links goods prices, that both countries remain diversified, and that producers still must meet price-equals-unit-cost conditions. In effect, trade has forced European workers to bear the burden of high unemployment to maintain not only their own high wage, but that in America as well.

How robust is this result? This is most easily illustrated by placing the same information in a framework of export supply and import demand. As we have seen, the minimum-wage constraint makes European export supply at price P^* multivalued, corresponding to various levels of unemployment. Graphically, this implies that the constrained export supply curve has a horizontal segment that corresponds to the

---

[7] This builds on a diagram of Brecher (1974a); see also Krugman (1995).

[8] An alternative way of seeing this is to note that $\beta(h^*; H^W, L^W)$ is linear homogeneous in H^W and L^W for a fixed h^*, and that opening to America is the same as doubling the world endowments (since Europe absorbs all of the unemployment).

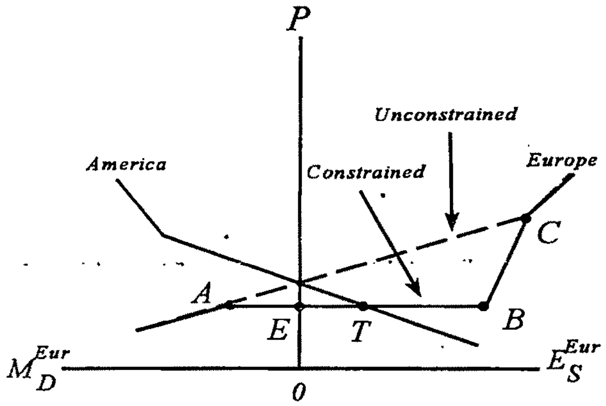


FIGURE 4. AMERICAN IMPORT DEMAND DETERMINES EUROPEAN UNEMPLOYMENT

Rybczynski segment of the constrained PPF. In Figure 4 this is depicted along with the American import demand curve for the case in which the countries are otherwise identical.

Now consider perturbations in the position of the American import demand curve, and consider their implications for unemployment and wages. So long as the American import demand curve cuts the export supply curve of Europe between *A* and *B*, the equilibrium price will be P^*, so the wage will be w^*. In all such cases, were Europe to abandon the minimum-wage policy, both European and American wages would fall in the free trade equilibrium. The effect of trade with America on European unemployment depends on whether the American import demand curve cuts the European export supply curve in the region *TB*, *ET*, or *AE*. If the equilibrium is in the region *TB*, then European unemployment will more than double, reaching its maximum at *B*. If in the region *ET*, European unemployment will be higher than in autarky, although at less than the double rate. If, instead, the equilibrium is in the region *AE*, then trade with America actually lowers European unemployment. It continues to be true in all three cases, though, that under free trade high American wages depend on the European minimum-wage policy fully as much as European wages do.[9]

III. Shocks from the South: Implications f Europe and America

The impact of growing trade with deve
ing countries on labor markets in develo
countries has been a source of great conc
It was at the heart of the vigorous opposi
of the AFL-CIO to the North American F
Trade Agreement, and has figured prominen
in the presidential campaigns of Ross P
and Patrick Buchanan. Correlate conce
have been raised over rapid integration
Eastern Europe into the markets of the Eu
pean Union.

The evidence for the growing trade, par
ularly in manufactures, is indisputable. In
period 1970–1990, the less-developed cou
tries (LDC) share in manufactured imports
the United States and Europe more th
doubled—rising to more than a third for
United States (Freeman, 1995). Correspon
ingly, the share of manufactures in total n
fuel exports from LDCs rose from less th
one-fourth in 1970 to nearly three-fourths
1989 (Wood, 1994).

However, what role—if any—this pl
in the factor market developments in Eur
and America is still in contest. Wood (19
has been the strongest proponent of the vi
that trade with the LDCs has mattered. Q
ified support for this view has come f
Edward E. Leamer (1995). By cont
Krugman and Robert Z. Lawrence (19
have argued forcefully that the volume of t
is insufficient to account for relative w
movements in the United States. Simil
Franco Modigliani (1995 p. 7) insists tha
argument that increased trade with develo
countries raised unemployment in Europ
"nonsense, reflecting economic illiteracy

Krugman (1995) examined the impact
trade shock on a minimum-wage Europe
a flexible-wage America, concluding th
would raise unemployment in the former
reduce wages in the latter. However his s

---

[9] When contemplating the recent unemployment experience of Europe, one should not read from this that it was *caused* by increasing integration with America. Rather this is designed to provide a simple window
trade-offs a country may face between a wage ta
unemployment in an open economy. In the be
case, the unemployment cost of any given wage
higher by a factor of two.

on a theoretical plane, belongs to the category earlier identified as "comparative." That is, it examines this one country at a time. As we will see, the results change importantly when we turn to a global approach, in which Europe and America exist within the same world economy.[10]

I begin again with a flexible-wage America, and a minimum-wage Europe. For simplicity, and without loss of generality, I shift the base case to consider that in which P^* is the American autarky price. This is also the European autarky price with the minimum wage in place. And so when free trade is allowed between America and Europe, P^* remains the equilibrium price, with no actual trade taking place. Europe has the same level of unemployment as in autarky, and both have the same wages in autarky as under free trade.

I now consider the introduction of a previously isolated region, which will be called the NICs, to the American and European free trade system. Assume that the NICs have a strictly downward-sloping import demand curve, and at prices P^* are importers of the good X. My concern will be to trace through the implications of this for Europe and America. In the diagram, this can be represented as a rightward shift of the joint American-NIC import demand curve. However, so long as Europe is not completely driven out of the Y industry, and given Europe's commitment to its minimum wage, the world price P^* does not change (see Figure 5).

At an unchanged world price, the American import demand is unchanged. So the United States absorbs none of the imports from the NICs. American wages continue to be protected by the European commitment to a high-wage policy. In Europe, by contrast, there will be a surge of imports from the NICs.[11] The impact of this is similar to that analyzed above. Given the European commitment to a high-wage policy, this will not show up in reduced wages, but rather in higher unemployment.

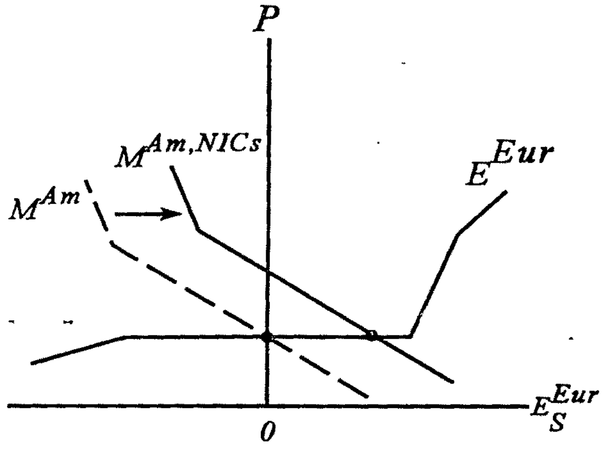


FIGURE 5. A NICS SHOCK RAISES EUROPEAN UNEMPLOYMENT, LEAVES AMERICA UNAFFECTED

This integrated global economy presents an important contrast with the results of Krugman (1995). There he treated the American and European cases as though they were part of different global economies. However, as is seen here, the links via goods trade are crucial for understanding the broader story. So long as Europe maintains a commitment to both free trade and a high-wage policy, America is fully insulated from the NIC shock.

IV. Global Labor Supplies, European Unemployment

A. Discussion

Leading labor economists have identified cross-country differences in the evolution of the labor force as an important explanatory factor in the divergent evolution of wages between continental Europe and the United

---

[10] This comparative theoretical approach is at work as well in an argument by the Center for Economic Policy Research (1995 p. 53). It is argued there that evidence, as in Lawrence and Matthew J. Slaughter (1993), exculpating Stolper-Samuelson in America, likewise refutes claims that trade with the newly industrialized countries (NICs) may have raised unemployment in Europe. However, as the discussion here makes clear, one reason that increased trade with the NICs may not account for falling wages in America is precisely that it *does* raise unemployment in Europe.

[11] The actual pattern of trade—whether NICs deliveries are to America or Europe—is inconsequential. The important point is that the price P^* defines the world net offer to Europe. Europe must absorb this net offer to maintain the price P^*, and so accept the employment consequences.

States and United Kingdom (Katz and Murphy, 1992; Katz et al., 1995). As insightful as this line of research has been, it would be more satisfying to consider this in a theoretical framework in which the cross-country structure of labor demand is endogenous. It might seem obvious that cross-country differences in the levels and growth rates of relative factor supplies should help to explain differences in relative wages. However, this need not be true when countries are linked by commodity trade. In fact, the absence of such a relation is the central message of Paul A. Samuelson's factor price equalization theorem (1949). Under the conditions in which that theorem is valid, the structure of relative wages depends on the evolution of the global labor force.[12] However cross-country differences in the composition or growth rates of the labor force will contribute nothing to an understanding of divergent wage trends. Given that it has been seen that the model with unemployment likewise yields factor price equalization, it can likewise be concluded that differential labor-force growth will not help to account for divergent wage patterns.

Although changes in relative factor supplies will not play a role in accounting for cross-country differences in relative wages, they will be essential for understanding the evolution of unemployment in Europe. There will be a correlate to the factor price equalization theorem: The *level* of European unemployment depends on global—not local—factor supplies. One consequence of this is that dramatic changes in American factor accumulation—the shifts due to baby booms and busts—may leave their most profound mark not on American but on European labor markets. By contrast, the unemployment *rate* in Europe will depend importantly on *where* factor accumulation occurs. The conditions under which European unemployment falls are less stringent when factor accumulation is in Europe th America.

The model also features an important as metry. The fixed European minimum wage sulates America from all shocks caused factor accumulation in Europe. But the rev is not true. Factor accumulation in Ame has very profound effects on Europe.

In the conventional Heckscher-Ohlin fr work, the central theorem dealing with end ment changes is that of Rybczynski. It that at fixed goods prices, an increase in endowment of a factor leads to a more-th proportional increase in output in the se that uses that factor intensively, and a con tion in the other sector. At the initial equi rium price this would be expected to creat excess supply of the good that uses that fa intensively, lowering its equilibrium rel price, and by Stolper-Samuelson reducing return to that factor. In the conventional ting, at fixed prices, changes in the endow of one country have no effect on output plies in the other.

B. *Derivation of Principal Results*

I now turn to a systematic evaluation of impact of factor accumulation in the pre model. I begin by considering the implica for the global economy, and then turn to amine how these effects are distributed cording to where the accumulation takes p It is assumed throughout that Europe's m mum wage at w^* remains binding and both economies are diversified.

Several observations will make the de tion of the results more transparent. Euro commitment to maintain the high wage likewise a commitment to maintain a dom relative price P^*, as given by the zero p conditions for the two goods. With free t this will also be the world relative pric saw above that the common goods prices factor price equalization. Cost-minimi firms with identical technologies will the the same factor intensities in both coun denoted h_X^* and h_Y^*. With P^* fixed in all e libria that are considered, these factor i sities will also be fixed. Because of ide and homothetic demand, P^* also fixes the in which X and Y will be consumed. Thi

---

[12] In fact, this extends to a more general setting. It is straightforward to write down a model in which only a subset of the world enjoys factor price equalization (Davis et al., 1997). Subject to the general restrictions underlying FPE (see Dixit and Norman, 1980), the site of endowment changes among those sharing FPE matters neither for comparative nor absolute wages.

[illegible] directly that world output at equilibrium must be supplied in the same proportion. This absorbs factors in the proportion h^* $[=\lambda^{-1}(P^*)]$. This represents both the minimum skill intensity that the world economy needs to yield the wage w^* in a flexible-wage world, and the actual skill abundance of total *employed* factors in the world with at least one economy with a minimum wage at w^*. Unemployment will arise when the actual world endowment, h^W, is such that $h^W < h^*$. The Rybczynski theorem is usually stated for a fixed exogenous price. Here we consider the case in which the price is held fixed by the commitment of Europe to maintain the minimum wage. Finally, recall that unemployment can be written as $U^* = L^W - H^W/h^*$.

Although my principal interest is in the effects of factor accumulation, it is convenient to begin with a version of the Heckscher-Ohlin theorem for trade between a flexible-wage economy and a minimum-wage economy. Let h^A denote the skill-to-labor ratio in flexible-wage America. Given a minimum wage w^* that binds where applied, and that both economies are diversified, the world employment ratio of skill to labor is fixed at h^*. The flexible-wage America will be an exporter of the skill intensive good if and only if $h^A > h^*$. The minimum-wage Europe will have a complementary trade pattern. The easiest way to see how the Heckscher-Ohlin theorem works here is simply to delete Europe's unemployed from Figure 2. They do not contribute income, so consume only from transfers that they will spend in the same pattern as those employed. Having done this, the diagram just becomes the conventional Heckscher-Ohlin model, although now in terms of *employed* factors.

I now turn to the results concerning accumulation of factors. A rise in the world endowment of skill raises employment of labor so as to maintain employed factors in the ratio h^*. As a result, output of both goods rises proportionally. If the skill accumulation occurs in flexible-wage America, it will have conventional Rybczynski effects, raising the output of X more than proportionally and decreasing the local output of Y. However this also affects Europe, which has a rise in employment. Recalling that $U^* = L^W - H^W/h^*$, it follows that $\Delta U^* = -\Delta H^W/h^*$. Thus it has Rybczynski effects skewed the other direction, with output of Y rising more than proportionally, and output of X declining. The net impact of the output changes in the two countries must, as noted, raise output of X and Y in proportion to the initial world output.

If instead the skill accumulation occurs in minimum-wage Europe, we have a very different picture. In Europe we fail to have conventional Rybczynski effects. Incipient excess supply of X threatens to raise the wage above w^* at the initial level of employment. This allows employment to rise by $\Delta N = \Delta H^W/h^*$. That is, total employment of factors in Europe has risen in exact proportion to total initial employment in the world as a whole, h^*. Thus, output in Europe expands proportional to initial world output. America is entirely unaffected, as its net trade vector is fully determined by the constant price ratio P^*.

Now consider the effect of accumulation of labor. At the global level, unemployment rises by the full increment to the labor force, so $\Delta U^* = \Delta L^W > 0$. There is no effect on global output. If the accumulation occurs in minimum-wage Europe, unemployment rises directly with no effect on output in either country. However, if accumulation of labor happens in America, there are very dramatic shifts. America continues to have conventional Rybczynski effects, raising the output of Y more than proportionally, and reducing the output of X. Income in America rises linearly, so $\Delta I^A = w^* \Delta L^A$. Europe is also strongly affected. In fact, European unemployment rises one for one with accumulation of labor in America. And European income falls dollar for dollar with the rise in American income. Its shift in output exactly offsets that of America, so that global output is unchanged.

V. South–North Migration and Unemployment

A. *Migration in a Model of Adjusted Factor Price Equalization*

In this section I develop a model of endogenous South–North migration that builds on Daniel Trefler (1993, 1996), and use this to consider the consequences for European unemployment. Think of a North composed of Europe and America. They are just as before, with Europe imposing a minimum wage and America having a flexible wage. Following Trefler (1993), it is assumed that factor price equalization between North and South holds, but only when labor is measured in efficiency (rather than natural) units. For simplicity I neglect the possibility that there are efficiency differences across countries in skill. Let L_E^W and N_E^W be the world labor force and employment when measured in efficiency units, and let $h_E \equiv H^W/N_E^W$. Then it is a straightforward application of our previous model to show that when Europe imposes a minimum wage at the level w^*, unemployment of labor is given as

$$U^* = L_E^W - H^W/h_E^*. \tag{6}$$

Consideration of the incentives for migration forces us to consider an issue familiar from the standard Ricardian trade model, namely the reason for variation in efficiency units per natural unit of labor. Cross-country variation that reflects only *individual characteristics* of the workers would provide no motive for migration, so are neglected.[13] If instead this also reflects cross-country *technological differences,* then labor will have a reason for migration—access to the superior Northern technology.[14] Let ϕ^N and ϕ^S reflect labor-augmenting productivity differences specific to the location of production. By choice of units, let $\phi^N = 1$. Assuming that labor productivity is higher in the North restricts $\phi^S \in (0, 1)$.

Let the efficiency equivalent labor force for the North and South be L_E^N and L_E^S respectively, where $L_E^N + L_E^S = L_E^W$. Let L^S be the number of nationals of the South, some of whom in the latter period move to the North. Let L^N be the number of nationals of the North (all of whom remain there). Then prior to
gration, the world labor force measured i
ficiency units is $L_E^W = L^N + \phi^S L^S$. W
migration of size M occurs from Sout
North, this changes to $L_E^{W\prime} = L^N + \phi^S(L$
$M) + M$. As labor was already in excess
ply, and noting that H^W and h_E^* are unalt
by the migration, the change in the world l
force measured in efficiency units is equi
lent to the rise in European unemploym
This is given as:

$$\Delta U^* = \Delta L_E^W = (1 - \phi^S)M. \tag{7}$$

Thus two parameters, ϕ^S and M, identify
contribution of South–North migration to E
ropean unemployment. In this world with F
in efficiency units, technologically based d
ferences in productivity (ϕ^S) will be identi
by the increment in wages to those who m
grate from South to North. M is the amoun
migration from developing countries to t
OECD.

B. *Migration Induced by Large Endowment Differences*

I now consider an alternative way to m
the impact of migration on unemployme
rigid-wage Europe. The most plausible a
native yields conclusions which are qua
tively similar, though quantitatively m
dramatic. Consider the case in which the
tive for migration is a failure of FPE bec
of large differences across countries in f
composition. Let there be three goods, X
and Z, in decreasing order of skill inten
As before, the minimum wage in the Nor
set in terms of the numeraire, Y. For sim
ity, assume the technologies are Leon
Skill-abundant countries of the North pro
the relatively skill-intensive goods X an
while the skill-scarce South produces Y a
As in the conventional two-factor, mu
production cone Heckscher-Ohlin model
labor-scarce North enjoys higher wages,
tivating immigration from the South. Con
the impact of the migration of a single w
from South to North. By the Rybczynski
orem, the migrant leads to an expansion
production in the North, and a contractio
X output at the initial prices. Moreove

---

[13] Moreover this would be reflected in a native-immigrant wage gap. Since empirical estimates suggest that such gaps are small (on the order of 10 percent) for workers with similar backgrounds, these can safely be ignored for present purposes.

[14] This also implies that "outsourcing" to the South may have the same implications as developed here for migration. The key element is the matching of Northern technology and Southern workers, which increases the effective world supply of unskilled labor, rather than the locale in which this matching occurs.

of an unskilled worker in the South leads an expansion of Y in the South (since this is a relatively skilled industry), and a contraction of Z output. Thus output movements in both countries are putting downward pressure on the relative price of Y (upward pressure on prices of both X and Z). This suggests the conventional result that migration places pressure for wages to rise in the source country and fall in the destination country. By how much would unemployment have to rise in the North to relieve this pressure? A full answer would require a more complete specification. Note, though, what would happen if unemployment in the North were to rise exactly one for one with immigration at the initial prices. Then the immigration would have left unskilled employment, so output in the North of X and Y, unchanged. However, there would still have been the effect on output in the South, raising output of Y and reducing that of Z. This suggests residual pressure for both the prices of X and Z to rise relative to Y to move toward equilibrium. If in fact the price of X did rise, this would push the Northern unskilled wage below the minimum. And this suggests that to meet the minimum wage, unemployment in the North would have to rise by a factor of *more* than one for one relative to migration from the South.

C. *An Empirical Application*

An extensive literature has considered the impact of immigration on labor-market outcomes of natives. Rachel M. Friedberg and Jennifer Hunt (1995) provide a valuable survey. The analytic approach pursued here is closest to that of George J. Borjas et al. (1992). I consider a simple counterfactual based on the theoretical models developed above. How would European unemployment have differed if unskilled immigration to the North (America and Europe) from 1970–1990 had been constrained to zero? The parsimonious specification of the two theoretical models above imply that two parameters suffice to answer this question: (1) the number of unskilled migrants from South to North; and (2) the wage gain for those who migrate. For the former, I will take a conservative definition, using the estimate of the 1970–1990 flows of unskilled migrants from the South who entered the labor force in the United States, a figure calculated as approximately 4.4 million (see the Appendix). For the latter I note that Trefler (1993), using International Labour Organization data that fit his FPE model well, calculated U.S. wages for aggregate labor relative to those of various other countries, for example coming up with figures ranging from a U.S. advantage of 20 to 1 relative to Bangladesh, 5 to 1 relative to Yugoslavia, and 3.3 to 1 relative to Colombia. No doubt an important element in this is differences in skill composition. However, direct evidence on the wage gains of individual migrants suggests that the gains that motivated the migration may be substantial. For the purposes of the calculations for Model A, I will assume conservatively that the wage gain relative to that earned premigration ranges between a factor of 2 and 3. Thus, under the hypothesis of the adjusted-FPE Model A, as reflected in equation (7), the counterfactual yields incremental European unemployment of between 2.2 million and 2.9 million. Under the hypothesis of Model B, the multicone approach, the corresponding incremental unemployment would be something greater than 4.4 million. If it is noted that in 1993 there were 17 million unemployed in Europe, the ceteris paribus figures suggest that absent this unskilled immigration, the same wage in Europe could have been targeted with a reduction of between one-eighth and one-quarter of this unemployment.

These figures are meant only to suggest an order of magnitude—that these effects could matter empirically. The data underlying these calculations is rough, while the model to which they are applied is sharp. Moreover, it would be inappropriate here, as in any ceteris paribus experiment, to infer from this that unskilled immigration "caused" the rise in European unemployment. Nevertheless, the possibility that unskilled immigration to the North could matter for European employment, given a commitment to a specific wage, is not implausible. This was in fact the point of the very restrictive immigration policies of Europe post-1973. Yet while the front door may have been shut, the open door provided for unskilled immigration to the United States implied incipient downward pressure on wages on both sides of the Atlantic. Europe's commitment to high wages forced it to adjust to these pressures via higher unemployment.

VI. Divergent Wage Experience

This paper began with the observation that American and European factor market experience has diverged in two respects: rising unemployment in Europe, but not in America, and a growing skilled-unskilled wage gap in America, but not in Europe. A great deal has been said about the former, and so the latter will be addressed here.

The conjunction of relative factor demand shocks and institutional differences has been an important theme in explanations of the divergent relative wage experience of America and Europe. Thus the demonstration that this wage divergence does not in fact arise in the standard model is an important analytic contribution of this paper. However, the divergent experience still needs to be understood.

From an analytic perspective the crucial insight is that shocks that move us from one integrated equilibrium to another will not explain the wage divergence. Thus endowment changes, demand shocks, and global technology shocks are all consistent with maintaining FPE across the countries, in spite of the institutional differences.

This leaves a variety of amendments to the baseline model that could potentially explain the divergent experience. One class concerns divergent evolution of policies or technological shocks. The factor price link comes through producers' zero profit conditions. For FPE we need free trade, zero internal taxes, common technologies, and diversified production, in addition to the conventional assumptions. Thus differences in the evolution of trade and fiscal policy, idiosyncratic technology shocks (e.g. convergence), or specialization can lead FPE to break down, so offer paths for interpreting the divergent wage experience. The role of technical change in understanding these factor market developments is considered at length in Davis (1998).

However one need not rely on divergent policies or technological shocks to explain the divergent wage experience. If one adds a little more structure to the basic framework, alternative accounts are possible. For exam [illegible] Davis and Trevor A. Reeve (1997) devel [illegible] model of endogenous human capital accu [illegible] lation in a framework that is otherwise [illegible] tical to that developed here. In their model [illegible] threat of unemployment has feedback ef [illegible] on individuals' decisions to accumulate [illegible] man capital. Since unemployment is prev [illegible] in Europe but not in America, even comm [illegible] shocks can lead to divergent accumulation [illegible] cisions, hence wage outcomes.

Thus the basic framework that has been [illegible] veloped is consistent with a variety of [illegible] counts of the divergent wage experie [illegible] between America and Europe. Import [illegible] the principal insights that have been develo [illegible] in this paper are robust to consideration of [illegible] of these variations.

VII. Conclusion

This paper has one overarching mess [illegible] Even when factor markets are strictly natio [illegible] with idiosyncratic institutional features, [illegible] cannot be considered in isolation when g [illegible] markets are global. This strongly suggest [illegible] importance of a unified "global" approach [illegible] explaining recent factor market developm [illegible] in Europe and America. This will be [illegible] important complement to existing stud [illegible] which are either of individual countries [illegible] comparative.

A striking example of the importance of [illegible] global perspective emerges in the contra [illegible] effects of trade on a flexible-wage Am [illegible] and a minimum-wage Europe. In the ca [illegible] example, a move from autarky to free [illegible] doubles the European unemployment [illegible] while leading American wages to conver [illegible] the high European level.[15]

---

[15] Why was European unemployment lower th [illegible] in America pre-1973, yet higher thereafter? One [illegible] would be to think of there being two compon [illegible] unemployment—one frictional and the other due to [illegible] ities. For a variety of reasons, America may have a [illegible] base frictional rate than Europe. If in one period the [illegible] ities are nonbinding in Europe, it may be found th [illegible] unemployment is relatively low. If in another perio [illegible] rigidities do bind, it may instead be found that une [illegible] ment there is relatively high.

Moreover, commodity trade leads local institutional features to have important spillover effects on other countries. These go far beyond the simple general equilibrium fact that they will affect the level of wages and the composition of production. They may fundamentally alter the nature of a country's relation with the international environment. For example, if Europe and America are both flexible-wage economies, then the entry of the NICs to world markets may depress wages in each. However, we saw that when Europe imposed a minimum wage, it absorbed the full impact of the NIC shock and wholly insulated America from its effects. That is, a local European institutional feature has sharply altered America's relation to the trading world.

The local institutional features may likewise introduce important asymmetries in the relations between the countries. As noted, America is wholly insulated from external shocks, including factor-supply shocks in Europe. By contrast, factor-supply shocks in America have very powerful effects in Europe. A striking example occurs with an increase in the supply of labor in America. This raises American income, yet lowers European income dollar for dollar, while raising European unemployment one for one.

The importance of considering the links via commodity trade is particularly important when thinking about the evolution of wages. In spite of important differences in factor market institutions, which in a closed economy would induce differences in factor prices, goods trade here insures full factor price equalization. Thus insofar as policy makers in a country see a trade-off between the two, the willingness of one to bear high unemployment spills over to raise wages in both.

The major analytics of this paper have been derived in a framework in which America and Europe replicate an integrated equilibrium with factor price equalization. This contributes greatly to the transparency of the results. As well it has emphasized the important point that Europe's commitment to a minimum wage cushions the impact of a variety of shocks on America. However it also suggests an important respect in which additional insight is indicated. A key stylized fact we want to understand is the divergent relative wage experience of America and Europe. Insofar as the shocks considered here move the world economy between equilibria featuring factor price equalization, they will not account for this fact. I have indicated a variety of directions in which the framework may be amended to account for this fact. This suggests the value of extensions that may help to identify which of these may matter most empirically.

This paper has derived these results in a stylized model. Care should be taken in reading these results too readily into actual historical experience. Nonetheless the issue that it raises of the powerful interaction between local factor market institutions and global goods markets is no doubt very important. And the results are sufficiently provocative to warrant closer examination.

APPENDIX: AUTHOR'S CALCULATION OF THE CONTRIBUTION OF SOUTH–NORTH MIGRATION TO THE WORLD EFFECTIVE STOCK OF UNSKILLED LABOR

As discussed in the text, over the last two decades Europe has pursued a restrictive immigration policy, while America has had strong growth in immigration, particularly among the unskilled. The stock of foreign population in Europe grew by 2.9 million from 1981–1991 (John Salt et al., 1994). However, much of this growth is from the end of the period, and naturally the growth of the unskilled labor force would be less than this. Since I want to be conservative, and since Europe's policy was overtly restrictive, for the purposes of my calculations I will assume that there was zero Southern unskilled immigration to Europe.

The calculations for America (here just the United States) must of necessity be imprecise. The desired category of immigrants is not readily available, so must be calculated. Here I detail those calculations, first for legal immigrants. The stock of foreign-born persons 18 years of age or older in the United States in 1990 was 17.7 million (U.S. Department of Commerce, Bureau of the Census [1990], Table 3: Social Characteristics of Selected Ancestry Groups by Nativity, Citizenship and Year of Entry: 1990). Of these, 6.9 million entered between 1980–1990 (U.S. Department of Commerce, Bureau of the Census, 1990). The 1970's saw approximately 4.5 million immigrants to the United States (Borjas, 1990 p. 6). Of these, close to 4 million would have been aged 18 or over (Borjas, 1990 p. 41). Summing, the 1970–1990 immigration of those aged 18 and over would have been approximately 10.9 million. The labor-force participation rate for male immigrants is approximately 90 per cent (Borjas, 1990 p. 41), while that for women is lower, which I will estimate by the native female participation rate of circa 50 percent. Thus the average labor-force participation rate will be taken as 70 percent. This implies that of the 10.9 million immigrants aged 18 and over, approximately 7.6 million would be in the labor force. Of the 1990 stock, approximately 60 percent of those aged 18 or over had a high-school education or less (U.S. Department of Commerce, Bureau of the Census, 1990). Thus of these 7.6 million in the labor force, a little over 4.5 million would fall into the category of the unskilled. Not all of these, though, come from the South. If we define the South operationally as Asia (except Japan), Africa, and the Americas (except Canada), approximately three-fourths of immigrants in the 1970's came from the South (Borjas, 1990 p. 230). If it is conservatively assumed that the unskilled were no more likely to come from the South than from the North, this would imply that the contribution of legal immigration from the South to the unskilled labor force in the North is approximately 3.4 million.

To this must be added the contribution to the labor force by illegal immigration. While newspaper reports of illegal immigration are often huge, the work of Borjas et al. (1991) suggests that in 1980 the stock of Mexican illegal immigrants to the United States was 1.8 million. I do not have data that provide similarly careful estimates for illegal flows for 1970–1990, or that incorporate the full complement of countries of the South. Realizing that this leaves out flows from the 1980's from Mexico, as well as all flows from other countries of the South, while it includes some who arrived prior to 1970, I will (with a great dollop of optimism) take the figure of 1.8 million as the relevant measure of inflows from 1970–1990. If it is assumed, as above, that approximately 90 per cent of these are aged and above, this reduces the figure to app mately 1.6 million. Assuming that the labo force participation rate again is 70 percent, implies that approximately 1.1 million are the labor force. It is known that educatio attainment of typical Mexican immigrants very low, with over three-fourths having le than a high-school education (Borjas et 1992). If those with a high-school educat are included among the unskilled, then will provide a figure of approximately 1 mi lion additional unskilled workers. If this summed to the contribution of legal immigr tion, a figure is arrived at which indicates South–North migration to the United Sta from 1970–1990 raised the labor force of unskilled by approximately 4.4 million wor ers. This estimate should be considered rough.

REFERENCES

Bertola, Giuseppe and Ichino, Andrea. "W Inequality and Unemployment: Unit States vs. Europe." *NBER macroeconom annual*, Vol. 10. Cambridge, MA: MI Press, 1995, pp. 13–54.

Boccard, Nicolas.; van Ypersele, Tanguy Wunsch, P. "Comparative Advantage, distribution and the Political Process: A spective on Social Dumping." Cente Operations Research and Economet (CORE) Discussion Paper No. 9651, Lo vain-la-Neuve, Belgium, 1996.

Borjas, George J. *Friends or strangers: The pact of immigrants on the U.S. econo* New York: Basic Books, 1990.

Borjas, George J.; Freeman, Richard B. and Lawrence F. "On the Labor Market Eff of Immigration and Trade," in Georg Borjas and Richard B. Freeman, eds., *migration and the work force*. Chic University of Chicago Press, 1992, pp. 44.

Borjas, George J.; Freeman, Richard B. and Kevin. "On the Labor Market Effects of migration and Trade," in John Abowd Richard B. Freeman, eds., *Immigra trade, and the labor market*. Chicago: versity of Chicago Press, 1991, pp. 77

Brecher, Richard A. "Minimum Wage Rates and the Pure Theory of International Trade." *Quarterly Journal of Economics*, February 1974a, *88*(1), pp. 98–116.

———. "Optimal Commercial Policy for a Minimum-Wage Economy." *Journal of International Economics*, May 1974b, *4*(2), pp. 139–49.

Center for Economic Policy Research (CEPR). *Unemployment: Choices for Europe*. London: CEPR, 1995.

Davis, Donald R. "Technology, Relative Wages, and Unemployment in a Global Economy." *European Economic Review*, 1998 (forthcoming).

Davis, Donald R. and Reeve, Trevor. "Human Capital, Unemployment, and Trade." Harvard Institute for Economic Research Working Paper No. 1804, Harvard University, August 1997.

Davis, Donald R.; Weinstein, David E.; Bradford, Scott C. and Shimpo, Kazushige. "Using International and Japanese Regional Data to Determine When the Factor Abundance Theory of Trade Works." *American Economic Review*, June 1997, *87*(3), pp. 421–46.

Dixit, Avinash K. and Norman, Victor F. *Theory of international trade*. Cambridge: Cambridge University Press, 1980.

Freeman, Richard B. "Are Your Wages Set in Beijing?" *Journal of Economic Perspectives*, Summer 1995, *9*(3), pp. 15–32.

Freeman, Richard B. and Katz, Lawrence F., eds. *Differences and changes in wage structures*. Chicago: University of Chicago Press, 1995.

Friedberg, Rachel M. and Hunt, Jennifer. "The Impact of Immigrants on Host Country Wages, Employment and Growth." *Journal of Economic Perspectives*, Spring 1995, *9*(2), pp. 23–44.

Helpman, Elhanan and Krugman, Paul R. *Market structure and foreign trade*. Cambridge, MA: MIT Press, 1985.

Katz, Lawrence F.; Loveman, Gary and Blanchflower, David. "A Comparison of Changes in the Structure of Wages in Four OECD Countries," in Richard B. Freeman and Lawrence F. Katz, eds., *Differences and changes in wage structures*. Chicago: University of Chicago Press, 1995, pp. 25–65.

Katz, Lawrence F. and Murphy, Kevin M. "Changes in Relative Wages, 1963–1987: Supply and Demand Factors." *Quarterly Journal of Economics*, February 1992, *107*(1), pp. 35–78.

Krugman, Paul R. "Europe Jobless, America Penniless?" *Foreign Policy*, Summer 1994, No. 95, pp. 19–34.

———. "Growing World Trade: Causes and Consequences." Mimeo, Stanford University, April 6–7, 1995.

———. "But For, As If, and So What? Thought Experiments on Trade and Factor Prices." Mimeo, Massachusetts Institute of Technology, November 1996.

Krugman, Paul R. and Lawrence, Robert Z. "Trade, Jobs, and Wages." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 4478, September 1993.

Lawrence, Robert Z. and Slaughter, Matthew J. "International Trade and American Wages in the 1980s: Giant Sucking Sound or Small Hiccup?" *Brookings Papers on Economic Activity*, 1993, (1), pp. 161–210.

Leamer, Edward E. "In Search of Stolper-Samuelson Effects Between Trade and U.S. Wages." Mimeo, Yale University and University of California, Los Angeles, 1995.

Matusz, Steven J. "International Trade, the Division of Labor, and Unemployment." *International Economic Review*, February 1996, *37*(1), pp. 71–84.

Modigliani, Franco. "The Shameful Rate of Unemployment in the EMS: Causes and Cures." Mimeo, Massachusetts Institute of Technology, 1995.

Murphy, Kevin M. "Comment—Wage Inequality and Unemployment: United States vs. Europe." *NBER macroeconomics annual*. Cambridge, MA: MIT Press, 1995, pp. 54–59.

Oslington, Paul. "Factor Price Equalization and Trade Patterns with Unemployment." Macquarie Economics Research Papers 2/96, Macquarie University, Australia, June 1996.

Salt, John; Singleton, Ann and Hogarth, Jennifer. *Europe's international migrants*. Lanham, MD: Bernan Associates, 1994.

Samuelson, Paul A. "International Factor-Price Equalisation Once Again." *Economic Journal*, June 1949, *59*(234), pp. 181–97.

Trefler, Daniel. "International Factor Price Differences: Leontief Was Right!" *Journal of Political Economy*, December 1993, *101*(6), pp. 961–87.

———. "Immigrants and Natives in General Equilibrium Trade Models." Mimeo, National Academy of Sciences, September 12, 1996.

U.S. Department of Commerce, Bureau of Census. *Statistical abstract of the Un States*. Washington, DC: U.S. Govern Printing Office, various issues.

Wood, Adrian. *North–South trade, empl ment, and inequality*. New York: Ox University Press, 1994.

Growth Cycles

By George W. Evans, Seppo Honkapohja, and Paul Romer*

We construct a rational-expectations model in which the economy switches stochastically between periods of low and high growth. When agents expect growth to be slow, the returns on investment are low and little investment takes place. But if agents expect fast growth, investment is high, returns are high, and growth is rapid. This expectational indeterminacy is induced by monopolistic competition and complementarity between different types of capital goods. Neither externalities nor increasing returns to scale are required. The equilibrium with growth cycles is stable under the dynamics implied by a simple learning rule. (*JEL* E32, O41)

One of the oldest conjectures in economics holds that self-fulfilling shifts in optimism or confidence can destabilize aggregate economic activity. Aggregate investment, which is notoriously volatile, has long been suspected of being susceptible to these fluctuations. In this paper, we construct an explicit model that exhibits this kind of expectational indeterminacy in investment. In a rational expectations equilibrium, changes in expectations cause the economy to cycle back and forth between periods with high levels of investment and rapid GDP growth and periods with low levels of investment and slow GDP growth.

The crucial assumption in our model is that different types of capital goods are complements. They are like personal computers, laser printers, and digital communication networks. If you are a producer, an increase in the number of complementary goods raises the demand for your good. If you are already in business, you expand output. If you are a potential producer who is not yet in business, now may be the time to enter. By increasing your output, you raise the demand for the complementary goods. We embed this self-reinforcing process in a standard model of monopolistic competition that characterizes the entry process for firms supplying new goods. Together, complementarity and monopolistic competition are all we need to generate expectational indeterminacy.

Other models have shown that multiple equilibria and expectational indeterminacy are possible in principle. (We review these models in Section III.) However, some of these models rely on assumptions that are implausible or address issues that are not central to macroeconomics. In what follows, we develop a model that is based on familiar and empirically plausible microfoundations. We do not rely on controversial assumptions. For example, no production relationship exhibits short-run increasing returns to scale, or equivalently, falling short-run marginal costs of production.[1] The only nonconvexity in the model is a fixed research and development cost that a firm must incur to introduce a new good.

---

* Evans: Department of Economics, University of Oregon, Eugene, OR 97403; Honkapohja: Department of Economics, University of Helsinki, FIN-00014, Helsinki, Finland; Romer: Graduate School of Business, Stanford University, Stanford, CA 94305. The authors are grateful for comments received in numerous seminars, workshops, and conferences. We received detailed comments from Paul Beaudry, Larry Christiano, Tim Cogley, Peter Howitt, and David Romer. We benefitted from support provided by the National Science Foundation, the program on Economic Growth of the Canadian Institute for Advanced Research, the SPES program of the European Union, and the Yrjö Jahnsson Foundation and the Academy of Finland, which gave grants to the Research Unit on Economic Structures and Growth in Helsinki.

[1] Susanto Basu and John G. Fernald (1997) provide evidence against short-run falling marginal costs. For a discussion of role of falling marginal costs in models with multiple equilibria, see Jess Benhabib and Roger Farmer (1994).

We eschew modeling assumptions that are problematic or distracting. For example, we avoid overlapping generations and assume that agents are infinitely lived. We specify the model entirely in real terms, so there is no fiat money. We assume that markets are complete, so agents can trade securities with payoffs that are contingent on the state of the economy at all dates. We do not assume that there are search-related externalities in the labor market. We do not allow for kinks in the demand curves that monopolists face.

Our analysis also goes beyond the mere demonstration that a rational expectations equilibrium with growth cycles exists. We show that the equilibrium is locally stable under a natural learning rule. As a result, arguments about stability under learning that were used to dismiss some of the implausible monetary equilibria cannot be used to rule out the behavior we identify. Moreover, the equilibrium is plausible in the sense that it makes only modest information processing demands on the agents who must form expectations.

To highlight the fundamental elements in the model, we present it in stages. In Section I, we start by describing the underlying perfect-foresight growth model with multiple equilibria and introduce the concept of stability under learning. Then, in Section II, we outline the full rational expectations model with stochastic shifts between the high- and low-growth states identified in Section I. A discussion of related literature is set forth in Section III; Section IV concludes.

I. The Perfect-Foresight Growth Model

A. *Modeling Strategy*

We cannot characterize the equilibrium in this model as the solution to a maximization problem. Nor can we linearize the model around a single steady state. Instead, we simplify the model so that the equilibrium can be characterized in terms of a small number of equations. The two most important simplifying assumptions are that the economy has a single state variable and that the preferences and technology are homogeneous. Together, these imply that at every date, the opportunities in the economy are the same except for a scale factor determined by the state va
This reduces the problem of calculating
finite path for dated consumption goods
stocks of capital to the simpler, finite di
sional problem of calculating the constan
portional amount by which consumption
the state variable grow.

Our preference structure is standard. We
sume that there is a representative consumer
maximizes the discounted expression for

$$\sum_{i=0}^{\infty} \beta^{t+i} U(C_{t+i}), \text{ where } U(C) = \frac{C^{1-}}{1-}$$

This utility function gives us the homoge
we need on the preference side of the
A consumer who is faced with a constan
terest rate r, will choose to have consum
grow at the constant rate g_C given by

$$(1) \qquad g_C = \frac{C_{t+1}}{C_t} = [\beta(1+r)]^{1/\sigma}.$$

When we introduce uncertainty later in the
the consumer will maximize expected utility.

We will characterize our equilibria in
of two equations in two endogenous varia
the interest rate and the rate of growth.
tion (1) gives us one of the two equations.
second comes from an arbitrage condi
from the technology side of the model.
cause our technology is complicated, we
ceed in steps, illustrating the relations
between r and g that emerge in simpler
and working up to the relationship for the
model.

B. *Linear Production*

The simplest production structure is
which output is linear in the stock of
and in which there is a one-for-one trad
between new capital and consumption:

$$(2) \qquad Y_t = BK_t,$$

$$(3) \qquad K_{t+1} - K_t = Y_t - C_t.$$

This is an example of what is now call
"*AK*" or linear growth model. (We
the symbol "*B*" instead of "*A*" to

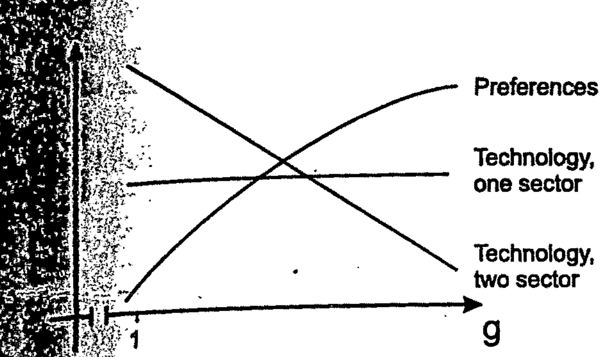


FIGURE 1. PERFECT-FORESIGHT EQUILIBRIA UNDER TWO DIFFERENT TECHNOLOGIES

confusion with our subsequent use of the symbol *A*.) The correct interpretation of the accumulation equation (3) is that consumption goods and capital goods are produced in different sectors that use the same production technology. In this case, the production possibility frontier (PPF) between additions to the capital stock and units of consumption goods is linear. The model can therefore be treated as if it has a single productive sector. By a suitable choice of units, the trade-off between capital and consumption goods can be taken to be one for one.

The behavior of this trivial model can be characterized by a graph that plots the equilibrium interest rate as a function of the equilibrium growth rate as in Figure 1. Equation (1) gives the upward-sloping curve labeled as the preference curve. The horizontal line $r = B$, which is labeled "technology, one sector" is determined by arbitrage on the production side of the model. Because of the homogeneity of the model, the values for r and g from the intersection of the technology and preferences curves determine everything about the equilibrium starting from any initial stock of capital [illegible]

Figure 1 also presents a second technology curve that is downward sloping. It characterizes the no-arbitrage condition when the production functions for consumption and capital goods are different. In this case, the production possibility frontier between additions to the stock of capital and consumption is nonlinear. We could assume a functional form for production in each sector and derive the expression for the production possibility frontier. Instead, we take a shortcut by assuming a simple functional form for the production possibility frontier itself. Let Y_t denote the maximum feasible amount of consumption goods that can be produced. Then output of consumption goods can be related to increases in the capital stock as follows:

$$(4) \qquad C_t = Y_t - K_t \chi\left(\frac{K_{t+1} - K_t}{K_t}\right).$$

In this expression, $\chi(\cdot)$ can be any increasing, convex cost function with $\chi(0) = 0$. This specification imposes the homogeneity of degree one in the state variable K that we need on the production side of the model. It reduces to equation (3) if χ is the identity function. The price of capital goods in units of consumption goods will be the derivative of consumption with respect to the stock of capital tomorrow. We will write this as $\chi'(g_K - 1)$, where $g_K = K_{t+1}/K_t$. One unit of consumption can now be invested to yield r units of additional consumption goods in each period. Alternatively, it can be used to purchase $\chi'(g_K - 1)^{-1}$ units of capital, which will yield a flow of B units of consumption goods in each period. Thus, a no-arbitrage condition on the technology side of the model imposes the equality $r = B/X'(g_K - 1)$. Because $X(\cdot)$ is convex, this expression yields the downward-sloping curve labeled "technology, two sector" in Figure 1.

Two features of this kind of model make this simple diagrammatic characterization of equilibria possible for both of these models. First, the accumulation equation for capital, (3) or (4), implies that if capital, output, and consumption grow at constant multiplicative rates, this rate must be the same for all three variables. The equilibrium can therefore be characterized using a common growth rate $g = g_C = g_K$. Second, because of the homogeneity of the specifications for preferences and the technology, the curves in the figure do not depend on the level of the state variable K. These two features will carry over to elaborations of this basic structure that we develop next. These extensions will allow for fixed costs of invention and complementarity between types of capital goods.

C. *Many Capital Goods*

To bring invention into the analysis, we start with a model of growth taken from Romer (1987). This model allows for the introduction of new types of capital goods, but its equilibrium behavior closely resembles the behavior of the one-sector model described above. Output of consumption goods is a function of labor (assumed in fixed supply) and the quantities of a large number of specialized capital goods:

$$(5) \qquad F(L, x(\cdot)) = L^{1-\alpha} \int_0^A x(i)^\alpha \, di.$$

Here $x(i)$ denotes the number of units of capital of type i that are in use. The upper limit A (assumed continuous for convenience) indicates the range of capital goods that have already been designed and can therefore be produced.

In addition to the production of final output, there are two other productive activities—inventing new types of capital goods and producing physical machines for each of the many existing types of capital goods. By analogy with the one-sector model described above, we can assume that the same production technology described by the functional $F(\cdot)$ can be used in all productive activities: (i) making consumption goods; (ii) making designs for new types of machines; and (iii) making physical machines for types that have already been designed. Consequently, the trade-off between the different types of output is linear. We can therefore write an expression for total output (or the maximum feasible output of consumption goods) as $Y_t = F(L, x(\cdot))$. Assume that it takes one unit of forgone consumption to make each additional unit of any machine and that it takes a units of forgone consumption to produce a design for a new type of machine. Ignoring depreciation for the moment, total output is split according to

$$Y_t = C_t + a(A_{t+1} - A_t) + K_{t+1} - K_t,$$

where K_t is defined as the total stock of machines measured in units of the cost of production for these machines:

$$K_t = \int_0^{A_t} x_t(i) \, di.$$

These two equations capture both the cost necessary before a new good can troduced and the constant marginal cost by a firm that supplies such a good after been introduced.

The market structure in this model i nopolistically competitive. Any firm th signs a new machine gets a patent o design. It produces machines and charge simple monopoly price for these machin equilibrium, the interest rate r adjusts so the present discounted value of the resu monopoly profits is just equal to the cost inventing a new type of machine. The profit condition in this equilibrium implies at every date, the ratio of A_t to K_t is co so we can work with a single aggregate variable $Z_t = aA_t + K_t$ that summarize effects of all previous investment decisi The accumulation equation can then be wr as $Y_t = C_t + Z_{t+1} - Z_t$.

In this model, the reduced-form expressio output Y is linear in Z. As a result, the inter poral zero profit condition is satisfied at a un value of the interest rate r. This value of independent of the state variable Z_t. (See Ro 1987, for details.) As a result, the analys growth is exactly as depicted in the single-se version of the technology in Figure 1. The duction technology determines a unique in rate. The preference specification then yiel rate of growth for the economy.

D. *Capital Goods That Are Complem*

As the introduction makes clear, our mai terest lies in capital goods that are comple with each other. To see how we can bring into the model, note that equation (5) us constant-elasticity-of-substitution aggrega summarize the effect on output of all the d ent types of capital goods. In (5) productio additively separable in different capital g To make them complements or substitutes need only introduce an additional expone the capital aggregate:

$$Y_t = L^{1-\alpha}\left(\int_0^{A_t} x_t(i)^\gamma \, di\right)^\phi.$$

Throughout the paper, we impose the restriction $\gamma\phi = \alpha$ to preserve homogeneity of degree 1. For values of ϕ greater than 1, the different capital goods $x_t(i)$ will be complements; an increase in the quantity of one raises the marginal product of the other.

In the additively separable model outlined above, it took a constant amount of forgone consumption to produce each design for a new type of capital good. If the capital goods are complements, this specification for the technology for producing designs would lead to explosive growth. Later designs would be more valuable because there are more complements around that raise the value of a new capital good. One simple way to offset the increasing incentives for discovering new goods is to assume that there is also a higher cost for designing goods with a higher index. Specifically, we will assume that it takes i^{ξ} units of forgone output to produce a design for good i. In equilibrium, the incentives for discovering new goods will just offset the costs of discovery if the following parameter restriction holds:

$$\xi = \frac{\phi - 1}{1 - \alpha}. \tag{6}$$

If this restriction does not hold, the model may still be well behaved, but will exhibit rates of growth that are locally increasing or decreasing over time. This would make it impossible to solve for a single, constant growth rate as we did in Figure 1 for the linear production model. To keep things simple, we impose the restriction (6) and look for equilibria exhibiting constant exponential growth.

In the extended model with complements, we can still write the accumulation equation for the single state variable as

$$Y_t = C_t + Z_{t+1} - Z_t.$$

The only difference arises in the slightly more complicated relationship between the stock of total capital Z and the other state variables. At date t, the expression for total capital Z is

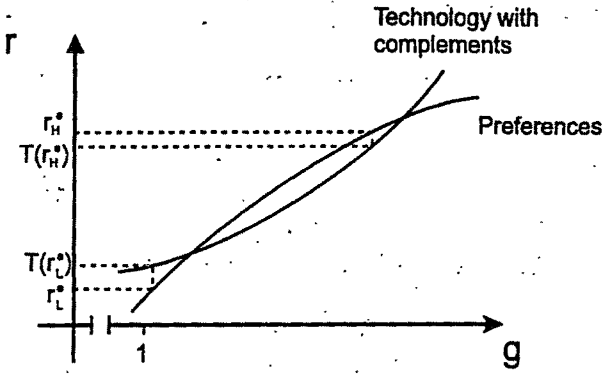


FIGURE 2. PERFECT-FORESIGHT EQUILIBRIA WITH COMPLEMENTARY INPUTS

$$Z_t = \int_0^{A_t} x_t(i)\, di + \int_0^{A_t} i^{\xi}\, di \tag{7}$$

$$= K_t + \int_0^{A_t} i^{\xi}\, di.$$

The first term, K_t, represents cumulative total investment in machines. (We will introduce depreciation later.) The second term describes cumulative total investment in the production of new designs.

The analysis of the perfect-foresight, monopolistically competitive equilibrium for this model proceeds exactly as it did in the model where the capital goods have additively separable effects on production. (Details are given in the Appendix.) The fundamental arbitrage result is derived from the intertemporal zero profit condition on investments in new designs. The key result that emerges from this analysis is that the no-arbitrage or zero profit relationship implied by the technology yields an upward-sloping locus of (r, g) pairs as illustrated in Figure 2. In this case, the growth rate g refers to the common growth rate for Y_t, C_t, and Z_t. Because designs are getting more costly, g_A is smaller than g. In fact, one can show that $g = g_A^{\xi+1}$. Because the preference relationship is also upward sloping, these curves may intersect more than once. For appropriate parameter values, they intersect twice, as illustrated in Figure 2. This means that the model has two distinct perfect-foresight equilibria. For any given initial value of Z_t, there is one perfect-foresight equilibrium in which the values for Y_t, C_t, and Z_t all grow at the high rate determined by the upper intersection. There is also a second perfect-foresight equilibrium in which they grow at the low rate determined by the lower intersection of the two curves.

The positive slope of the technology locus is fundamental to all of the results that follow. The intuition behind this slope is easy to understand. Because of the complementarity between the different capital goods, a firm that invents a new type of good today will face a demand for its good that increases with the quantities and varieties of other investment goods that will be introduced tomorrow. Start at an interest rate and a growth rate that yield zero intertemporal profit and increase the growth rate for the economy. At a faster rate of growth for other goods, the present discounted value of the stream of future profits for a new invention will be higher. Competition for resources by potential new entrants will therefore drive up the interest rate until the present discounted value of the future profits is driven back down to the cost of inventing a new good.

E. *Stability Under Learning*

The relatively simple model characterized in Figure 2 is the basis for a model of macroeconomic fluctuations in which changes in expectations cause the economy to switch between the high- and low-growth states. However, it leaves two important issues unresolved. First, we must let agents take account of the possibility of such switches when they form their expectations. We will deal explicitly with this issue in Section II.

The second issue suggested by Figure 2 is stability. Whenever we see two equilibria in an economic model, we have come to expect that one of them will be stable and that the other will be unstable. The goal in this section is to identify the precise sense in which one can speak of an equilibrium point as being stable or unstable. Then in subsection F of this section, we will modify the model so that it has three perfect-foresight equilibria, two on the ends that are stable and one in the middle that is unstable.

It does not make sense to think of these equilibria as being stable or unstable in the usual sense—that from a large class of initial values for the state variable the equilibrium dynamics take the economy away from some steady-state points and toward others. The standard analysis does not apply to our model because it has a single state variable Z and the dynamic equations that determine the growth rate do not depend on the level of Z. For any initial value of Z, the economy can select either of the perfect-foresight paths characterized by the intersections in the Figure 2.

We can, however, inquire into the stability properties of the equilibria by going outside of the perfect-foresight model and asking what the dynamics would be if agents had to learn about the equilibrium values of variables by observing the behavior of the economy. If plausible learning dynamics lead the agents away from some perfect-foresight equilibria and toward other equilibria, we can use this difference to distinguish between more- and less-plausible perfect-foresight equilibria. That is, we can use stability under the learning dynamics as a selection criterion when there are multiple perfect-foresight equilibria.

To conduct the learning analysis in a model such as this, one must first extend the model to a temporary equilibrium framework. To do so, we use a simple scheme in which households base their actions on an expected interest rate. Fortunately, this analysis can be carried out using the same curves that characterize the equilibrium in (r, g) space. Given an expected interest rate r^e, households choose their savings to generate the rate of growth of wealth and consumption implied by (1), the equation that determines the preference curve in Figure 2. Firms take this growth rate of consumption and the implied, and equal, growth rate of total capital Z as given and project it into the future. At this expected growth rate for the economy there is a unique realized interest rate that is consistent with no-arbitrage on the technology side of the market. This is the value of r that is determined by inserting the given growth rate into the technology curve in the figure. If we take the composition of these two operations, we have the following mapping from expected interest rates to realized interest rates:

$$r = T(r^e). \tag{8}$$

By construction, the interest rates corresponding to perfect-foresight, balanced growth paths are fixed points of T.

After consumers observe the realized rates, they adjust their expectations about future interest rates. For example, if the interest rate that consumers expected was lower than the interest rate that is realized, they will revise their forecasts of interest rates upward. Out of equilibrium, both firms and consumers make mistakes but the process will converge over time to the stable equilibrium if it starts from nearby values for r and g. As the economy converges, these mistakes become smaller. In the limit, the rules lead to correct forecasts by both firms and consumers at constant equilibrium interest rates and growth rates.

To generate an explicit dynamics for the interest rate, we use an adaptive learning scheme,

$$r^e_{t+1} = r^e_t + \delta_t(r_t - r^e_t), \tag{9}$$

where $\delta_t = \delta/t$. The sequence $\{\delta_t\}$ is known as the gain sequence. It determines the extent of the adjustment of expectations to forecast errors. A decreasing sequence $\{\delta_t\}$ implies that consumers become less responsive to forecast errors as more data accumulates. For example, with $\delta_t = 1/t$ and appropriate initial values, equation (9) implies that r^e_{t+1} is equal to the average of past values of r_t. Together, equation (9) and the mapping $r_t = T(r^e_t)$ define a dynamic system which can be analyzed in the vicinity of a fixed point. If we linearize this system around an equilibrium r, we can show that the fixed point is locally stable if $T'(r) < 1$.[2]

The two equilibria in Figure 2 do indeed have different stability properties. The function T is the composition of the two functions graphed in the figure. At points where they intersect, the size of T' relative to 1 is determined by the relative slopes of the two curves. A simple geometrical argument shows that it is the lower equilibrium that is stable under learning. In Figure 2, r^e_L denotes an expected interest rate that is below the interest rate in the low-growth perfect-foresight equilibrium and r^e_H denotes an expected interest rate that is also below the interest rate in the high-growth equilibrium. As the figure shows, when the mapping T is applied to r^e_L, it leads to a value $T(r^e_L)$ that is above r^e_L, so adjustment moves the expected interest rate in the right direction. In contrast, when T is applied to r^e_H, it leads to a value that is below r^e_H, so the learning dynamics drives the economy away from the high-growth equilibrium.[3]

If one accepts, as we do, the notion that any perfect-foresight equilibrium that is unstable under simple learning rules is unlikely to be realized in practice, one is forced to conclude that the kind of multiplicity exhibited in Figures 2 is unlikely to have any practical relevance. To have a candidate for a model with multiple stable perfect-foresight equilibria, we will need a model that generates more intersections of the two curves in this figure.

F. *Multiple Stable Perfect-Foresight Equilibria*

Careful readers will no doubt have guessed where the argument is now headed. We showed in the analysis of Figure 1 that a two-sector structure that generates a standard, nonlinear trade-off between the production of investment goods and consumption goods tends to make the technology curve slope down in (r, g) space. We have just shown that complementarity between different types of

---

[2] This kind of learning rule is particularly appropriate in an extension of this model which allows for small stochastic shocks, but we do not introduce these shocks explicitly into the analysis here. Because δ_t goes to zero there is no risk of instability in the adjustment process for large negative values of $T'(r)$. For a discussion of these kinds of learning algorithms, see Evans and Honkapohja ([illegible]b).

[3] Note that there are alternative ways of formulating the temporary equilibrium and corresponding learning rules. For example, we could start with firms that have an expected growth rate g^e_t for the economy. The arbitrage condition on the technology side of the model will lead to an interest rate r. Given this interest rate, consumers would use the preference curve to decide how much to save. Firms would then adjust their expected growth rates when they see the realized rate of growth implied by these savings decisions. It can be easily verified that a steady state is stable under (8) and (9) if and only if it is stable under this alternative scheme.

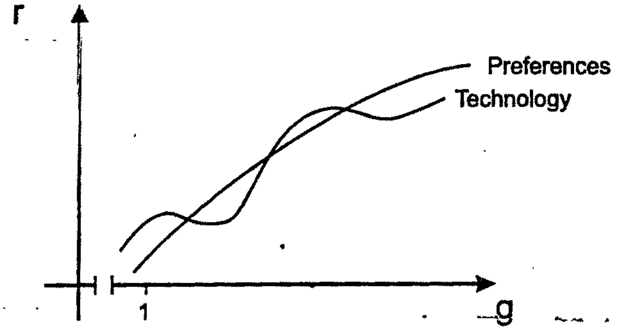


FIGURE 3. PERFECT-FORESIGHT EQUILIBRIA WITH COMPLEMENTS AND A TWO-SECTOR TECHNOLOGY

capital goods tends to make the technology curve slope up. If we combine these two elements in the same model, we should be able to make the curve have increasing and decreasing segments, so it can intersect the preference curve arbitrarily often.

Specifically, we now adopt the specification for the production possibility frontier between consumption and investment that we used previously. Thus, the accumulation equation for Z takes the form

$$C_t = Y_t - Z_t\chi\left(\frac{Z_{t+1} - Z_t - D_t}{Z_t}\right). \tag{10}$$

As in the previous specification, $\chi(\cdot)$ is a convex cost function. In this final expression for the model, we have also included a term D_t that takes account of any depreciation of the physical capital stock. As in the previous model with complements, the total stock of Z is related to the stock of designs and machines by equation (7).

This specification implies that the same production function can be used to produce machines and inventions. More precisely, it can be used to produce the raw input that is used to make the inventions. (Recall that it takes more input to produce an invention with a higher index.) This specification lets us aggregate A and K into a single state variable Z, which substantially simplifies the analysis. Equation (10) implies that this common production function for the durable inputs is different from the production function for making consumption goods. Finally, we assume that physical machines depreciate at the exponential rate d but that design blueprints do not depreciate, so D_t can be expressed as $D_t = dK_t$. In all other respects, the model is the same as the previous model with complementarities.

TABLE 1—STABLE PERFECT-FORESIGHT EQUILIBRIA

Growth rate g	Interest rate r	Price of capital
1.002	0.04	1.1
1.049	0.05	2.5

Figure 3 illustrates one type of simple equilibrium that we have generated using this combined structure. The downward-sloping portions of the technology curve represent points where the curvature in the production possibility frontier between consumption goods and investment goods is high so its effects overwhelm the effects of complementarity. The upward-sloping portions of the technology curve represent regions where the production possibility frontier is relatively flat so the effects of complementarity dominate. By the analysis of subsection E of this section, the middle equilibrium will be unstable under learning. The low- and high-growth equilibria will be stable under learning.

Table 1 reports values of the growth rate and the interest rate in the high- and low-growth equilibria in the basic parameterization of the model. (Numerical values for the parameters are given in the Appendix, subsection A.) By construction, the interest rates in these two equilibria were arbitrarily set at 4 percent and 5 percent, respectively. The corresponding net growth rates for Z, C, and Y are 0.2 percent per year and 4.9 percent per year. The table also reports the values for p^z, the price of the investment goods relative to consumption goods. In the high-growth state, where a larger fraction of GDP is devoted to investment, investment goods are more expensive to produce on the margin, so their prices are higher. Even though investment goods are more expensive in the high-growth state, a higher level of investment can be sustained because of the positive reinforcing effects that result from the complementarities between the different types of investment goods.

Figure 4 illustrates the nature of the implied production possibility frontier between Z

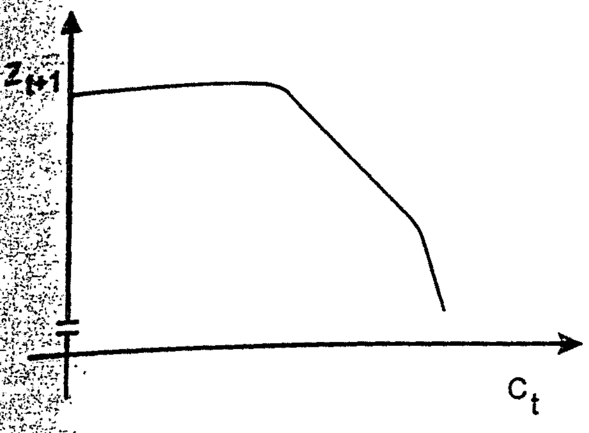


FIGURE 4. PRODUCTION POSSIBILITY FRONTIER BETWEEN CONSUMPTION AND NEXT PERIOD'S STOCK OF TOTAL CAPITAL

and C_t for a given level of Z_t. The two regions of sharp curvature correspond to the two downward-sloping regions of the technology curve in Figure 3. This picture also makes it clear that we are not relying on any assumptions about short-run increasing returns in the production of new investment goods. As noted in the introduction, some models generate multiple growth rates by introducing short-run increasing returns that make investment goods cheaper when a larger fraction of total output is devoted to investment. We generate multiple growth rates *despite* the fact that investment goods are more expensive when investment is higher and growth is more rapid. We reiterate that the only nonconvexity in the model is the fixed cost associated with invention.

II. Growth with Expectational Indeterminacy

A. *Equilibrium Fluctuations*

Using the building blocks of the perfect-foresight analysis we now construct a complete model of fluctuations between high- and low-growth states. In an equilibrium with fluctuations, consumers and firms must take account of these switches. Formally, we can do this by introducing a random variable s_t that matters only because agents use it to forecast the behavior of the economy. This kind of "extrinsic uncertainty" is susceptible to various interpretations. For a discussion, see for example the chapter on "Market Psychology" in Costas Azariadis (1993). We assume that s_t has no direct effect on preferences, technology, or government policy, but in an extension of the model, it could have a small direct effect as well as a larger indirect effect acting through expectations.

In the "high" state for s_t, firms and households (correctly) anticipate higher investment and more rapid growth during the coming period. However, they also take full account of the probability of a switch to the low-growth state in future periods. Formally, we let s_t take values from $\{1, 2\}$ and assume that it follows a two-state Markov chain with transition probabilities $0 < \Pi_{ij} < 1$ for $i, j = 1, 2$. These transition probabilities are exogenously given parameters.

Formally, all of the traditional elements in the specification of our model—the preferences and the technology—remain the same after we introduce s. All that changes is that, in our calculation of an equilibrium, we look for solutions to the equations generated by the model in a larger solution space in which consumers and firms can make decisions that are contingent on the random variable s. The perfect-foresight equilibria we have already calculated remain equilibria in which consumers and firms make the same choices for the two different values of s. If we find that there are other equilibria in which output and prices vary with the realization of s, we will say that the equilibrium exhibits expectational indeterminacy. This, we think, is a more suggestive description of the behavior in question than the conventional label of a "sunspot equilibrium."

The timing of the realization of s and of decisions is as follows. In period $t - 1$, total potential output is realized. Consumers make decisions about how much to consume and how much to save. Given the net-of-depreciation stocks of machines $(1 - d)K_{t-1}$ and designs A_{t-1}, their decisions about how much to save will determine the total stock of capital Z_t that will be carried over into period t. At the beginning of period t, the value of s_t is realized. Firms then make their decisions about production contingent on the observed value of s. In particular, they decide how to split the new investment between additions to K and increases in A. (In the equilibria we

TABLE 2—GROWTH CYCLE WITH 5-PERCENT PROBABILITY OF SWITCHING STATES

State s_t	Growth rate: $g_Z(s)$	Interest rate: $r(s)$	Price of capital: $p^z(s)$	Consumption/Output: $C/Y(s)$
s = low	1.0083	0.0383	1.435	0.978
s = high	1.0437	0.0518	1.565	0.881

consider, firms will always want to add positive amounts to the stocks of both A and K.) When s signals fast growth, firms will allocate more of their investment to increases in A because a design is more valuable when the economy is growing rapidly. Given these decisions by firms, total output will be realized at the end of period t. Consumers observe the state s_t and decide how much to consume and how much to save. In the high-growth and high-interest-rate state, they will save more and consume less. (This positive response of savings to increases in the interest rate follows from our assumption that the parameter σ in the utility function is less than one.) Thus, Z will tend to increase by a larger factor during high-growth states. Undepreciated capital and the saving decision of households then determine the quantity of capital Z_{t+1} which is carried into $t + 1$ and must once again be divided between increases in K and A.

This is the minimal extension of our perfect-foresight framework in which we can consider the possibility of expectational indeterminacy, but it nevertheless introduces a substantial degree of additional complexity. In particular, the variable s_t introduces a second state variable into the analysis, and the system can no longer be reduced in any meaningful way to a two-equation, two-variable graphical analysis. (The equations for this system are given in the Appendix.)

One way to study the behavior of a system of equations is to look for numerical solutions. We present one such solution based on the same parameter values that we used in calculating the two stable perfect-foresight equilibria in Section I. The only difference here is that we allow for a 5-percent probability that between any two periods, the economy switches from the high-growth to the low-growth state or vice versa, i.e., $\Pi_{11} = \Pi_{22} =$ 0.95. Table 2 provides the description of such a growth cycle that we calculated numerically.

In this equilibrium with expectational indeterminacy, K and A always grow between periods t and $t + 1$, regardless of the values taken on by s_t and s_{t+1}. This means that we do not have to consider the issues that would arise if firms wanted to disinvest, convert K into A, or convert A into K.[4] In the two states, the consumption-income ratio C/Y varies as one would expect, given our assumption that saving responds positively to increases in returns. On the firm side, investment behavior would be dominated by firm perceptions about the state of aggregate demand and the rate of growth in the near term. Investment by any individual firm would go up when interest rates are higher and growth is more rapid. Finally, note the relatively small amount of variation in the price of capital goods p^z compared to the markedly different values that it takes on in the two different perfect-foresight equilibria. This reflects the changes in behavior that forward-looking agents make when they anticipate future transitions between high- and low-growth states.

B. *Existence of Stable Growth Cycles*

It is no accident that the growth cycle that we calculated numerically generates growth and interest rates that are close to the values that arise in the two perfect-foresight equilibria we constructed in Section I. The underlying reason is that a pair of perfect-foresight equilibria can also be thought of as an equilibrium in which decisions depend on the realizatio of s but in which the transition probabilities between the two states are zero. Because the system of equations depends smoothly on the transition probabilities, the implicit function theorem tells us that for transition probabilities near zero, the behavior of the economy in the high- or low-growth state will be close to the behavior of the economy in the corresponding perfect-foresight equilibrium. Moreover, if the individual perfect-foresight equilibria are stable under learning, then the full equilibrium with expectational indeterminacy that switches between states that are near these equilibria will also be stable under learning.

---

[4] Formally, our computed equilibrium would not be affected by adding the restrictions $A_{t+1} \geq A_t$ and $K_{t+1} \geq (1 - d)K_t$.

It is easiest to state the existence result and the convergence result at the same time. To do so, we must first describe what the learning dynamic is in our full equilibrium with expectational indeterminacy. For the case of perfect foresight, we formulated a model of learning based on a single variable, the expected interest rate. We must now extend this approach to the full model in which variables can depend on s_t.

The equations that determine the full equilibrium can be interpreted as a mapping that takes a pair of expected interest rates (r_1^e, r_2^e) to a pair of realized interest rates (r_1, r_2):

$$(11) \qquad (r_1, r_2) = \tilde{T}(r_1^e, r_2^e).$$

(The equations themselves are presented in the Appendix.) The equilibrium with expectational indeterminacy is a fixed point of this mapping with values $r_1 \neq r_2$. Because it is a fixed point, expectations are correct. As above, households make savings decisions on the assumption that interest rates r_1^e and r_2^e will prevail in the two states and producers expect consumers to persist with the savings behavior this induces. Then the production side of the market will clear at the interest rates r_1 and r_2 in the two different states. That is, r_1 and r_2 will be consistent with zero expected profit for an entrant into the invention of new goods. In a hypothetical dynamic determined by the learning process, the time t realized interest rate r_t would be given by

$$(12) \qquad r_t = \tilde{T}_i(r_{1,t}^e, r_{2,t}^e) \qquad \text{if } s_t = i, \qquad \text{for } i = 1, 2,$$

where $\tilde{T}_i$ denotes the ith component of the mapping $\tilde{T}$.

We adopt the following simple state-contingent adaptive learning scheme. In state s, agents will compare the interest rate they expected to obtain in this state with the interest rate that actually obtains. They will then adjust their expectation of interest rates in this state in the direction of the realized interest rate. As before, the amount of the adjustment depends on the amount of data that they have already collected:

$$(13) \qquad r_{i,t+1}^e = r_{i,t}^e + \frac{\delta}{N_i(t)}(r_t - r_{i,t}^e), \qquad \text{if } s_t = i, \quad \text{and}$$

$$r_{i,t+1}^e = r_{i,t}^e \qquad \text{if } s_t \neq i.$$

In this expression, $N_i(t)$ represents the number of times the variable s_t has taken value i up until time t. Equations (12) and (13) define a dynamic system in the two variables r_1^e, r_2^e, driven by the exogenous stochastic variable s_t.

An equilibrium growth cycle, i.e., a rational expectations equilibrium with expectational indeterminacy, is a fixed point (r_1, r_2) of (11) in which $r_1 \neq r_2$. When we want to be explicit, we can express these equilibrium interest rates as functions of the underlying transition probabilities, $r_1(\Pi)$ and $r_2(\Pi)$. With this notation, we can describe conditions under which a nondegenerate growth cycle exists and which has the property that for nearby initial conditions, the sequences of state-dependent interest rates generated by the learning dynamics converge to the state-dependent interest rates that obtain in equilibrium.

PROPOSITION 1: *Suppose that the model outlined in Section I has multiple distinct, perfect-foresight, balanced growth equilibria. Let the mapping T be as defined in that section, so that the interest rates corresponding to two of these equilibria satisfy* $r_L = T(r_L)$ *and* $r_H = T(r_H)$. *Assume without loss of generality that* $r_L < r_H$. *Suppose that these two values both satisfy the perfect-foresight stability criterion,* $T'(r_L) < 1$ *and* $T'(r_H) < 1$. *Then there exists a neighborhood of the* 2×2 *identity matrix I with the property that for any transition* *probability matrix* Π *in this neighborhood, there is an equilibrium growth cycle* $(r_1(\Pi), r_2(\Pi))$ near (r_L, r_H) *which is locally stable under the learning dynamics* (12) *and* (13).

The proof of the proposition, together with a precise statement of the phrase "locally stable under learning," is given in the Appendix.

Finally, we note that we have verified directly that the numerical example presented in the last section is locally stable in the natural sense. In fact, we used a version of this method to calculate the equilibrium growth cycle itself.

III. Discussion of Related Literature

Models with multiple equilibria are potentially of interest for understanding both macroeconomic fluctuations and economic development.[5] Our focus here is only on the first of these issues. There are many strands in the literature on aggregate models with some form of indeterminacy or multiplicity of equilibria. We now review how the model of this paper fits into this extensive literature.

Some of the first models of multiple equilibria in macroeconomics were derived from the analysis of fiat money in William A. Brock (1975). Subsequently, Guillermo A. Calvo (1978) and David Cass and Karl Shell (1983) allowed for overlapping generations of agents. Peter Diamond (1982) introduced search externalities in the labor market. Michael Woodford (1991) exploited kinked demand curves that could arise under imperfect competition. As macroeconomists, we believe that the mechanisms identified in these papers are not central for generating expectational indeterminacy in real economies. However, as economic theorists, we have been able to take advantage of the insights that have developed in the exploration of these models. To cite just one example, our analysis of stability under learning builds on results developed by Peter Howitt and R. Preston McAfee (1992) in the context of a model with search externalities in the labor market.

In the more recent literature, one strand (surveyed in Roger Guesnerie and Woodford, 1992) focuses on the possibility of endogenous fluctuations in dynamic economies populated with overlapping generations of agents. Another strand has stressed complementarities in the decision-making of different agents that arise from externalities or imperfect competition. These complementarities can lead to multiple equilibria, sometimes even to a continuum of equilibria. Recent surveys include Joaquim Silvestre (1993) and Matsuyama (1995). Many of these models do not, however, explicitly construct rational expectations models with endogenous fluctuations arising from expectational indeterminacy. Some of this work was motivated by macroeconomic fluctuations, but a large and growing fraction of these models arises from the analysis of long-run growth. We will return to a discussion of the growth models below.

Our model is descended from two lines of work. One, illustrated by Andrei Shleifer (1986), shows how aggregate fluctuations can arise in a model where innovation is motivated by monopoly rents. The second line, developed by John Bryant (1983), emphasizes the importance of complementarity between intermediate inputs in production.[6] Recent work on complementarities and imperfect competition has used models with unique equilibria with large multipliers, e.g., Satyajit Chatterjee and Cooper (1993), or models with expectational indeterminacy, e.g., Chatterjee et al. (1993). The formal methods used in these papers are different, but the substantive economic assumptions are qualitatively the same. We have chosen the second approach because it locates the source for economic fluctuations squarely on expectations.

The unexpected discovery that growth models sometimes exhibit various forms of indeterminacy or multiplicity of equilibria lent new impetus to work on these issues.[7] These papers generate the underlying multiplicity of equilibria by demonstrating the existence of a continuum of perfect-foresight equilibria converging to a single steady state or of sunspot solutions in a neighborhood of this steady state. This, for example, is the main thrust of Benhabib et al. (1994) that uses a model very similar to ours. In general, this literature relies on local linearizations, whereas our analysis depends fundamentally on global nonlinearities.

---

[5] See Alwyn Young (1993) and the survey by Kiminori Matsuyama (1995) for a discussion of the relevance of models with multiple equilibria and multiple steady states to the development experience. Benhabib and Jordi Galí (1995) also use these kinds of models to characterize the behavior of different developing countries.

[6] See also Russell W. Cooper and Andrew John (19 for an influential and clear discussion of complementarities in macroeconomic models.

Models with a continuum of equilibria approaching a single steady state can offer a novel strategy for justifying the stickiness of prices and the power of monetary policy. (See Farmer, 1991; Paul Beaudry and Michael Devereux, 1993, for models of this type.) However, a potential difficulty with solutions in the neighborhood of an indeterminate steady state is that they may not be robust under learning. For example, in the simplest model in this genre (the overlapping generations model with money) Robert E. Lucas, Jr. (1986) found that adaptive learning schemes selected the determinate rather than the indeterminate steady state. Evans and Honkapohja (1994) confirm this lack of robustness for a somewhat more general class of models.

Another type of indeterminacy in growth models, described in the survey by Benhabib and Galí (1995), results from the presence of a finite number of distinct steady states.[8] For some initial values of the state variables in these models, there may be a finite number of discrete equilibrium paths, each of which leads to a distinct steady state. These models have steady states in levels of per capita income with no growth or in which all steady states exhibit growth at the same rate.

Our approach is similar in the sense that we start from an underlying model with a finite number of distinct equilibria rather than a continuum of equilibria. However, in our case, these distinct equilibria have different growth rates rather than different levels. In our model, fast and slow growth are possible starting from any initial condition. Like Richard Startz (1994), we adopt this two-state approach because of empirical work showing that a two-state model with a high-growth and a low-growth state gives a reasonably good description of the variation in U.S. economic growth (James D. Hamilton, 1989). Jean Pierre Drugeon and Bertrand Wigniolle (1996) have also developed a model of endogenous growth with Markovian fluctuations in growth rates, but they, like Startz, rely on the assumption of large, positive technological externalities that increasingly seem to be empirically implausible.

IV. Conclusions

The most difficult challenge for macroeconomists is to identify plausible mechanisms that produce large, coördinated aggregate fluctuations. A related, but easier, challenge is to find mechanisms that generate persistence in economic downturns. Our model of growth cycles, based on complementarities and monopolistic competition, does both.

As we emphasize in the previous section, our analysis builds on a variety of theoretical results about multiple equilibria and expectations. Our aim is to show that these phenomena can arise in a plausible model that makes sensible predictions. The model suggests that when firms make investment decisions, they pay careful attention to expectations about future economic activity and understand that the expectations of other firms will determine what takes place. Announcements about monetary policy are one example of a variable that could coordinate the expectations that different firms form, but there are many candidate variables that can play this role.

---

[7] For recent work in this area, see Christophe Chamley (1993), Benhabib and Farmer (1994), Benhabib and Roberto Perli (1994), Benhabib et al. (1994), Farmer and Jang Ting Guo (1994), Galí (1994), and Danyang Xie (1994). For general discussions, see Timothy Kehoe et al. (1992), Benhabib and Aldo Rustichini (1994), and Benhabib and Galí (1995). The economic assumptions required for local indeterminacy are sometimes controversial. For example, Benhabib and Farmer (1994) and Farmer and Guo (1994) require increasing marginal returns to labor. (See Stephanie Schmitt-Grohé, 1997, for a discussion of the assumptions required in various models.) In other cases the fit with the data is poor in dimensions we regard as crucial. For example, Galí (1994) obtains expectational indeterminacy only if the procyclicality of the investment share is

[8] These can arise, for example, because of threshold effects as in Azariadis and Allan Drazen (1990), or because of various kinds of increasing returns. Again, see the discussion in Matsuyama (1995). Galí (1995) presents an example in which multiple steady states arise because of an inverse relationship between markups and the aggregate capital stock.

To summarize the contribution that our paper makes relative to the existing literature, it is useful to separate the formal differences in our modeling strategy from the substantive differences in our economic assumptions. At a formal level, our model generates persistent growth rather than convergence to a steady state. The underlying growth model exhibits a finite number of distinct equilibria, whereas many other models rely on a continuum of equilibria. We focus not on indeterminacy per se, but rather on a pattern of stochastic switches between fast and slow growth. We explicitly show that our candidate equilibrium survives even when we introduce learning. We avoid increasing marginal productivity of inputs (or falling marginal cost of output). We do not allow for external effects.

The crucial economic assumptions in our growth model are selected because of their plausibility: fixed start-up costs, *ex post* monopoly profits, and complementarities between types of capital goods. These elements yield a model of economic fluctuations driven by self-fulfilling shifts in expectations about the returns to investment. Expectations also enter in a way that we find plausible: the strength of business confidence depends on short- and medium-term forecasts of the rate at which future sales and profits are expected to expand in the entire economy.

The kind of theoretical exercise undertaken here cannot prove anything about the quantitative importance of complementarity or market power. However, it can show that these effects can generate self-reinforcing pressures. It can also show that nothing in standard economic theory rules out the possibility that these pressures are large enough to generate booms and busts. Once this theoretical point is clear, economists may be encouraged to look at the data from a slightly different perspective. When they do, they may find that there is some basis for the old conjecture that "animal spirits" can lead to fluctuations in aggregate investment.

APPENDIX

A. *Perfect-Foresight Equilibria*

We will derive the general expression for the no-arbitrage condition depicted in Figure 3. The monotonically increasing version depicted in Figure 2 emerges as a special case.

Final goods producers are price takers in the market for capital goods. In equilibrium they equate the rental rate with the marginal product of capital

$$\text{(A1)}\quad R_t(j) = \frac{\partial Y}{\partial x_t(j)} = L^{1-\alpha}\left(\int_0^{A_t} x_t(i)^{\gamma}\,di\right)^{\phi-1} \times \phi\gamma x_t(j)^{\gamma-1}.$$

The firms that supply capital goods act as monopolistic competitors. Formally, they purchase general-purpose capital Z and convert it into specialized types of capital goods at the rate of one unit of general-purpose capital for one unit of specialized differentiated capital. Our timing convention implies that production is realized at the end of a period. At the end of period t, a capital-goods supplier will earn revenue $R_t(j)x_t(j)$. In the beginning of period t, it takes $x_t(j)$ units of Z to produce the $x_t(j)$ units of capital of type j that this firm owns and rents out. The rental cost on these units of Z is $r_t x_t(j) p_t^z$, where p_t^z is the price in units of consumption of one unit of Z. A fraction d of the capital goods depreciates in period t, so the total cost to the capital-goods supplier in period t is $(r_t + d)x_t(j)p_t^z$. This firm selects its output to maximize profits, taking as given the inverse demand curve for the rental price in equation

(A1) and the price p^z. This leads to the markup rule

$$\text{(A2)}\quad R_t(j) = (r_t + d)p_t^z/\gamma \equiv R(r_t, p_t^z).$$

Since this holds for each intermediate good j, we have by symmetry, that the provision of intermediate goods is at the same level for all types j and we denote their common value at time t by $x_t(j) = x_t$. It follows that aggregate output at time t is given by

$$\text{(A3)}\quad Y_t = L^{1-\alpha}(x_t^\gamma A_t)^\phi,$$

where

$$\text{(A4)}\quad x_t = LA_t^{(\phi-1)/(1-\alpha)}\left[\frac{R_t}{\phi\gamma}\right]^{-1/(1-\alpha)},$$

with R_t given by (A2). Profits (at the end of period t) for each type of intermediate good are given by

$$\begin{aligned}\text{(A5)}\quad \pi_t &= R_t x_t - (r_t + d)p_t^z x_t\\ &= (1-\gamma)R_t x_t\\ &= \Omega A_t^{(\phi-1)/(1-\alpha)}((r_t + d)p_t^z)^{\alpha/(\alpha-1)},\end{aligned}$$

where $\Omega = \phi^{1/(1-\alpha)}(1-\gamma)\gamma^{(1+\alpha)/(1-\alpha)}L$.

To produce a particular type of intermediate good j, a firm must first incur the cost of inventing and designing this good. Recall that the cost of a new invention for good j is j^ξ units of investment goods Z where $\xi > 0$. Because the firms are monopolistic competitors, at each time t the quantity of designs is determined by a zero profit condition which states that the fixed cost of the last design created at time t must be just equal to the present discounted value of the stream of monopoly rents it offers. In a perfect-foresight equilibrium with balanced growth, the interest rate is constant and the required zero profit condition for the marginal good created at time t (which will have an index A_t) is given by

$$p_t^z A_t^\xi = \sum_{s=0}^{\infty} \pi_{t+s}(1 + r)^{-(s+1)}.$$

This expression for discounted profits reflects our assumption that the cost of inventing a good is incurred at the start of period t but the first profit is not realized until the end of the period. Under balanced growth, the rate of growth of designs $A_{t+1}/A_t = g_A$ will be constant. Substituting (A5) into this sum gives

$$p^z A_t^\xi = A_t^{(\phi-1)/(1-\alpha)}\Omega((r + d)p^z)^{-\alpha/(1-\alpha)} \times \sum_{s=0}^{\infty} (g_A^\xi)^s(1 + r)^{-(s+1)}.$$

Because we assume that $\xi = (\phi - 1)/(1 - \alpha)$, we can cancel A_t^ξ and sum the series to obtain the zero profit condition

$$\begin{aligned}\text{(A6)}\quad g_A^\xi &= 1 + r\\ &\quad - \Omega(p^z)^{-1/(1-\alpha)}(r + d)^{-\alpha/(1-\alpha)}.\end{aligned}$$

Gross investment is equal to the net increase in Z minus a correction for the depreciation of physical capital goods. Let η represent the fraction of Z that is allocated to physical capital goods. We derive an expression for η below. Production of general-purpose capital Z takes place in a competitive sector, so that the price of capital goods in terms of consumption is

$$p^z = -\frac{dC_t}{dZ_{t+1}} = \chi'\left(\frac{Z_{t+1}}{Z_t} - 1 + d\eta_t\right).$$

[Recall that the function $\chi(\cdot)$ specified in (10) gives the trade-off between consumption and Z.]

At any date t, the following relationship must hold between Z, x, and A:

$$\text{(A7)}\quad Z_t = x_t A_t + \int_0^{A_t} i^\xi \, di = x_t A_t + \frac{A_t^{1+\xi}}{1 + \xi}.$$

Along a balanced growth path, the output-capital ratio Y/Z, the growth rate of capital, the interest rate, and p^z are constant. Using

(A4) to substitute for x_t, one obtains an expression for Z_t in terms of A_t:

$$\text{(A8)}\quad Z_t = A_t^{1+\xi}\left\{(1+\xi)^{-1} + L\left[\frac{R_t}{\phi\gamma}\right]^{1/(\alpha-1)}\right\} \equiv A_t^{1+\xi}\mu(R_t).$$

Note that we have defined $\mu(R)$ as a shorthand expression for the term in braces in (A8). Since $\eta_t = A_t x_t / Z_t$, it follows from (A7) and (A8) that

$$\eta(R_t) = 1 - \frac{1}{(1+\xi)\mu(R_t)}.$$

Thus we can express the price of general-purpose capital as

$$\text{(A9)}\quad p^z = \chi'(g_Z - 1 + d\eta(R(r,p^z))).$$

Finally, from the above expression (A8) for Z_t it follows that $g_Z = g_A^{1+\xi}$. Substituting in the expression (A6) for g_A and using $\xi = (\phi - 1)/(1-\alpha)$ we have

$$\text{(A10)}\quad g_Z = [1 + r - \Omega(p^z)^{-1/(1-\alpha)} \times (r+d)^{-\alpha/(1-\alpha)}]^{(\phi-\alpha)/(\phi-1)}.$$

Together, equations (A9) and (A10) give an implicit definition of the locus of pairs (g, r) that are consistent with the zero profit condition. In the special case where χ is the identity, p^z is a constant that can be taken to be equal to 1 and equation (A10) alone determines this locus. These loci are the "technology" curves in Figures 2 and 3.

To complete the analysis of the perfect-foresight equilibrium, it remains to verify that consumption and the stock of Z grow at the same rate. Note first that if we substitute (A4) into (A3), this generates an expression for Y_t in terms of A_t. Using the expression for Z in terms of A from equation (A8), it follows that the ratio of output to total capital is constant along a balanced growth path:

$$\text{(A11)}\quad \frac{Y_t}{Z_t} = \frac{L\left[\frac{R_t}{\phi\gamma}\right]^{\alpha/(\alpha-1)}}{(1+\xi)^{-1} + L\left[\frac{R_t}{\phi\gamma}\right]^{1/(\alpha-1)}} \equiv \lambda(R_t).$$

Next, we solve for the optimal pattern of consumption by the method of undetermined coefficients. We guess that consumption takes the form $C_t = \theta Z_t$ and solve for θ. From (1) we obtain

$$\text{(A12)}\quad g_Z = [\beta(1+r)]^{1/\sigma}.$$

Finally, we note that we can use equation (10) to conclude that

$$\text{(A13)}\quad \theta(r,p^z) = \lambda(R(r,p^z)) - \chi\{[\beta(1+r)]^{1/\sigma} - 1 + d\eta(R(r,p^z))\}.$$

It remains for us to specify the parameters we used in the calculation of our numerical example and the precise form we selected for the function χ. In the calculation of the two stable perfect-foresight equilibria, we used the following parameter values: $\alpha = 0.4$, $\Omega = .015$, $\phi = 4$, $d = 0.15$, $\beta = .962$, $\sigma = 0.21$, $L = 0.356244889$.

A continuous, piecewise linear form $\tilde{\chi}$ underlying the χ function was constructed as follows: $\tilde{\chi}(0) = 0$, $\tilde{\chi}'(y) = p_0$ for $y < y_1$, $\tilde{\chi}'(y) = p_1$ for $y_1 \leq y \leq y_2$, and $\tilde{\chi}'(y) = p_2$ for $y > y_2$. The smoothed function sets $\chi(y) = \tilde{\chi}(y)$ for $y \leq y_1 - \delta_1$, $y_1 + \delta_1 < y \leq y_2 - \delta_2$, and $y > y_2 + \delta_2$. In the interval $(y_1 - \delta_1, y_1 + \delta_1]$ $\chi(y)$ is the quadratic which satisfies $\chi(y_1 - \delta_1) = \tilde{\chi}(y_1 - \delta_1)$, $\chi'(y_1 - \delta_1) = \tilde{\chi}'(y_1 - \delta_1)$. Similarly, in the interval $(y_2 - \delta_2, y_2 + \delta_2]$ $\chi(y)$ is another quadratic satisfying $\chi(y_2 - \delta_2) = \tilde{\chi}(y_2 - \delta_2)$, $\chi'(y_2 - \delta_2) = \tilde{\chi}'(y_2 - \delta_2)$. For the numerical example we selected the values $p_0 = 0.555406$, $p_1 = 1.46523$, $p_2 = 5.06686$, $y_1 = 0.0199733$, $y_2 = 0.0535351$, $\delta_1 = 0.001$, $\delta_2 = 0.0005$.

At these parameter values nonnegativity of consumption is satisfied. We must also verify that discounted utility is finite along all feasible growth paths and that the transversality condition at infinity is satisfied for the consumer. To establish both of these results, it is sufficient to show that the maximum feasible rate of growth satisfies $g_{\max}^{1-\sigma} < 1/\beta$. To do this, we impose an additional constraint on the function χ which limits the maximum feasible rate of growth of Z. Specifically, we assume that $\chi'(y)$ goes to infinity for values of y above a threshold $y_{\max}$. In our numerical example, the high growth rate for Z satisfies the inequality $g_z < 1 + r = \beta^{-1}g_z^{\sigma}$, so we know that we can pick a value of $y_{\max}$ that is strictly larger than the value of y that obtains in the high-growth state. It follows that the imposition of this upper bound will not change the values of the perfect-foresight equilibria that we have calculated without imposing this bound.

B. *Equilibria with Expectational Indeterminacy*

Once we introduce the random variable s, both households and firms face uncertainty about the state of the world in the subsequent period. We must therefore introduce state-contingent prices for output. Let $q(s_t, s_{t+1})$ denote the price at the end of time t of one unit of output available at the beginning of period $t + 1$, contingent on the realization of s_{t+1}. Thus $q(s_t, s_{t+1})$ prices only the uncertainty associated with the realization of s_{t+1}. Formally, we assume that no clock time elapses during this interval. The pricing of goods over time is associated with a state-contingent interest rate $r(s_t)$ which denotes the sure net one-period interest rate on consumption good loans over the interval extending from the beginning of period t to the end of period t. Thus the q's price only the uncertainty and r's price only the time lag. Since one unit of output in state s_t at the end of period t necessarily is one unit of output in either state one or two at the beginning of period $t + 1$, we have the basic arbitrage formulae

$$q(i, 1) + q(i, 2) = 1 \qquad \text{for } i = 1, 2. \tag{A14}$$

We need to restate all of the equations for the consumer and firms in terms that allow for the uncertainty associated with s. We begin with the arbitrage relationship for firms. The problem for the producer of the final good continues to be a static, one-period problem, so we can write $R_t(j) = \partial Y/\partial x_t(j)$ just as we did before. This producer considers only the rental cost R_t of the specialized capital goods that it uses. In contrast, the firms that produce, own, and rent out these specialized capital goods now must worry about capital gains or losses on their goods that may be induced by changes in the price of general purpose capital, p^z. When the economy switches to a slow-growth equilibrium, the rate of investment falls. This makes the cost of new capital goods fall, which makes the value of all existing capital goods fall. We can price the uncertainty facing the producer of a specialized type of capital using $q(s_t, s_{t+1})$ and the price p^z. End-of-period profits for the firm supplying the specialized intermediate capital good j are

$$\begin{aligned}
&[R_t(j)(s_t)x_t(j) \\
&\quad + (1 - d)p^z(1)x_t(j)]q(s_t, 1) \\
&\quad + [R_t(j)(s_t)x_t(j) \\
&\quad + (1 - d)p^z(2)x_t(j)]q(s_t, 2) \\
&\quad - (1 + r(s_t))p^z(s_t)x_t(j).
\end{aligned}$$

It is convenient to introduce the notation

$$\begin{aligned}
\hat{p}^z(s_t) &= p^z(1)q(s_t, 1) \\
&\quad + p^z(2)q(s_t, 2) \qquad \text{for } s_t = 1, 2
\end{aligned} \tag{A15}$$

for the "expected" price of capital. Maximizing profits then leads to the equation

$$\begin{aligned}
R(s_t) &= \gamma^{-1}[(1 + r(s_t))p^z(s_t) \\
&\quad - (1 - d)\hat{p}^z(s_t)] \qquad \text{for } s_t = 1, 2.
\end{aligned} \tag{A16}$$

As before, $R_t(j) = R_t$ and $x_t(j) = x_t$ turn out not to depend on j and (A4) becomes $x_t(s_t) = LA_t^{(\phi-1)/(1-\alpha)}[R_t(s_t)/\phi\gamma]^{-1/(1-\alpha)}$.

End-of-period profits π_t are given by

$$\pi(s_t, A_t) = A_t^{(\phi-1)/(1-\alpha)}\Gamma(R(s_t), \hat{p}^z(s_t), r(s_t), p^z(s_t)),$$

where

$$\Gamma(R, \hat{p}^z, r, p^z) \equiv L\left[\phi\gamma\left(\frac{R}{\phi\gamma}\right)^{\alpha/(\alpha-1)} + \left(\frac{R}{\phi\gamma}\right)^{1/(\alpha-1)} \times [(1-d)\hat{p}^z - (1+r)p^z]\right].$$

The decisions to introduce designs are also dynamic and made under uncertainty. It is most convenient to approach the problem in terms of asset pricing. Let $P(s_t, A_t)$ denote the price of one unit of designs in period t contingent on the state s_t and the aggregate quantity of designs A_t. After they have been introduced, all designs are equivalent, so we can use a single price for all of them. Since each unit of design capital yields current profits $\pi(s_t, A_t)$, we have the arbitrage (or zero profit) condition

$$P(s_t, A_t) = \frac{q(s_t, 1)}{1 + r(s_t)}(\pi(s_t, A_t) + P(1, A_{t+1})) + \frac{q(s_t, 2)}{1 + r(s_t)}(\pi(s_t, A_t) + P(2, A_{t+1})).$$

At every date, the price of a unit of design capital is determined by the cost of producing the last design,

$$P(s_t, A_t) = A_t^{\xi} p^z(s_t).$$

Writing $g_A(s_t, s_{t+1}) = A_{t+1}(s_{t+1})/A_t(s_t)$ for the rate of growth of design capital and combining the above equations we obtain the equations:

$$\text{(A17)}\quad p^z(i) = (1 + r(i))^{-1} \times [\Gamma(R(i), \hat{p}^z(i), r(i), p^z(i))] + (1 + r(i))^{-1} \times [q(i, 1)p^z(1)g_A(i, 1)^{\xi} + q(i, 2)p^z(2)g_A(i, 2)^{\xi}]$$

for $i = 1, 2$.

We look for a solution in which θ and g_Z functions of the current state i of s_t. The ra of total investment to investment in designs given by

$$(1 + \xi)(Z_t/A_t^{1+\xi})(s_t) = (1 + \xi)\mu(R(s_t))$$

where μ is defined by (A8). From the prec ing equation the four possible growth ra $A_{t+1}/A_t = g_A(s_t, s_{t+1})$ are related to capi growth by

$$\text{(A18)}\quad g_A(i, j) = \left(g_Z(i)\frac{\mu(R(i))}{\mu(R(j))}\right)^{(1+}$$

for $i, j = 1,$

Turning to the household, there are now ler equations for the two state-contingent sets which pay one unit of consumption at 1 in a specified state. For expected sta contingent interest rates $r^e(s_{t+1})$ these fir order equations are

$$q(s_t, s_{t+1})\left[\frac{C_{t+1}}{C_t}(s_t, s_{t+1})\right]^{\sigma} = \beta\Pi(s_t, s_{t+1})(1 + r^e(s_{t+1})).$$

Here $\Pi(s_t, s_{t+1})$ is the transition probabil from state s_t to s_{t+1}. We will also use notation $\Pi_{ij} = \Pi(i, j)$, $i, j = 1, 2$. The timal solution for the household allows consumption to capital ratio to depend on state s_t:

$$(C_t/Z_t)(s_t) = \theta(s_t).$$

Combining these last two equations yields

$$\text{(A19)}\quad \beta\Pi_{ij}(1 + r^e(j)) = q(i,j)\left[\frac{\theta(j)}{\theta(i)} g_Z(i)\right]^{\sigma} \qquad \text{for } i, j = 1, 2.$$

Additional restrictions on the values of $\theta(i)$ come from the product transformation function (10) which holds state by state. Thus we have a state-contingent version of equation (A13):

$$\text{(A20)}\quad \theta(i) = \lambda(R(i)) - \chi[g_Z(i) - 1 + d\eta(R(i))] \qquad \text{for } i = 1, 2.$$

Finally, we have the state-contingent version of the capital price equation (A9)

$$\text{(A21)}\quad \hat{p}^z(i) = \chi'(g_Z(i) - 1 + d\eta(R(i))) \qquad \text{for } i = 1, 2,$$

and the rational expectations condition

$$\text{(A22)}\quad r(i) = r^e(i) \qquad \text{for } i = 1, 2.$$

The equations (A14)–(A22) form a system of 22 nonlinear equations in the 22 unknowns $r(i), r^e(i), R(i), q(i,j), \theta(i), p^z(i), \hat{p}^z(i), g_A(i,j)$, and $g_Z(i)$. These characterize an equilibrium with expectational indeterminacy. Omitting (A22), these equations implicitly determine the mapping $(r(1), r(2)) = \tilde{T}(r^e(1), r^e(2))$ described in the main text.

C. Existence and Stability of Growth Cycles

For notational convenience write $r_i = r(i)$ and $r_i^e = r^e(i)$ for $i = 1, 2$. We now sketch the proof of the proposition given in the text and show that generically there exist growth cycles for probabilities in a neighborhood of the identity matrix when there are distinct perfect-foresight, balanced growth paths. Let r_L, r_H denote the interest rates for two distinct balanced growth paths. By assumption $r_L \neq r_H$. Define a new function

$$W((r_1, r_2), (\Pi_{11}, \Pi_{22})) \equiv \tilde{T}((r_1, r_2), (\Pi_{11}, \Pi_{22})) - (r_1, r_2),$$

where we have made the dependence of the map $\tilde{T}(\cdot)$ on the probabilities Π_{11}, Π_{22} explicit. If we introduce vector notation $\mathbf{r} = (r_1, r_2)$ and $\mathbf{\Pi} = (\Pi_{11}, \Pi_{22})$, we can write the key condition of the implicit function theorem as the requirement that the matrix $\mathbf{D_r W}$, evaluated at that pair of steady states, is nonsingular.

By inspecting equations (A14), (A15), (A16), (A17), (A18), (A19), (A20), and (A21), it is seen that for the first component of $\mathbf{W}$ we have

$$\text{(A23)}\quad W_1((r_1, r_2), (1, \Pi_{22})) = T(r_1) - r_1 \qquad \text{for } \forall r_1, r_2, \Pi_{22},$$

where $T(r)$ is given by equation (8). To show this we first note that $\Pi_{11} = 1$ implies $q(1, 1) = 1$. Substituting $q(1, 1) = 1$ into the relevant members of equations (A19) yields the preference curve (A12). Second, inserting $q(1, 1) = 1$ and $q(1, 2) = 0$ into (A15), (A16), and (A17), and taking into account (A18), we obtain the technology curve (A10). Finally, it is easily seen that (A21), (A15), and (A16) yield (A9). A similar argument can also be made that $W_2((r_1, r_2), (\Pi_{11}, 1)) = T(r_2) - r_2$ when $\Pi_{22} = 1$.

From the above one can compute

$$\text{(A24)}\quad \mathbf{D_r W}((r_L, r_H), (1, 1)) = \begin{pmatrix} T'(r_L) - 1 & 0 \\ 0 & T'(r_H) - 1 \end{pmatrix}$$

at (r_L, r_H). Thus, growth cycles exist in a neighborhood of (r_L, r_H) provided $T'(r_L) \neq 1$ and $T'(r_H) \neq 1$.

Next, we consider the stability of an equilibrium under adaptive learning. Evans and Honkapohja (1994) show that the dynamics defined by (13) are locally stable—i.e., the trajectories starting from any point in a neighborhood of the equilibrium converge almost surely to the equilibrium—whenever a condition known as E-stability is satisfied.[9] For E-stability it is required that the eigenvalues of $\mathbf{D}_r\mathbf{T}$, evaluated at an equilibrium growth cycle, have real parts less than one. By continuity of eigenvalues and (A24), this condition is met in a neighborhood of (r_L, r_H) provided $T'(r_L) < 1$ and $T'(r_H) < 1$. This proves Proposition 1.

REFERENCES

Azariadis, Costas. *Intertemporal macroeconomics.* Oxford: Blackwell, 1993.

Azariadis, Costas and Drazen, Allan. "Threshold Externalities in Economic Development." *Quarterly Journal of Economics,* May 1990, *105*(2), pp. 501–26.

Basu, Susanto and Fernald, John G. "Returns to Scale in U.S. Production: Estimates and Implications." *Journal of Political Economy,* April 1997, *105*(2), pp. 249–79.

Beaudry, Paul and Devereux, Michael. "Monopolistic Competition, Price Setting, and the Effects of Real and Monetary Shocks." Mimeo, University of British Columbia, 1993.

Benhabib, Jess and Farmer, Roger. "Indeterminacy and Increasing Returns." *Journal of Economic Theory,* June 1994, *63*(1), pp. 19–41.

Benhabib, Jess and Galí, Jordi. "On Growth and Indeterminacy: Some Theory and Evidence." *Carnegie-Rochester Conference Series on Public Policy,* December 1995, *43,* pp. 163–211.

Benhabib, Jess and Perli, Roberto. "Uniqueness and Indeterminacy: On the Dynamics of Endogenous Growth." *Journal of Economic Theory,* June 1994, *63*(1), pp. 113–42.

Benhabib, Jess; Perli, Roberto and Xie, Danyang. "Monopolistic Competition, Indeterminacy and Growth." *Ricerche Economiche,* December 1994, *48*(4), pp. 279–98.

Benhabib, Jess and Rustichini, Aldo. "Introduction to the Symposium on Growth, Fluctuations and Sunspots: Confronting the Data." *Journal of Economic Theory,* June 1994, *63*(1), pp. 1–18.

Brock, William A. "A Simple Perfect Foresight Monetary Model." *Journal of Monetary Economics,* April 1975, *1*(2), pp. 133–5

Bryant, John. "A Simple Rational Expectations Keynes-Type Model." *Quarterly Journal of Economics,* August 1983, *98*(2), pp. 525–28.

Calvo, Guillermo A. "On the Indeterminacy of Interest Rates and Wages with Perfect Foresight." *Journal of Economic Theory,* December 1978, *19*(2), pp. 321–37.

Cass, David and Shell, Karl. "Do Sunspots Matter?" *Journal of Political Economy,* April 1983, *91*(2), pp. 193–227.

Chamley, Christophe. "Externalities and Dynamics in Models of Learning by Doing." *International Economic Review,* August 1993, *34*(2), pp. 1–27.

Chatterjee, Satyajit and Cooper, Russell W. "Entry and Exit, Product Variety and the Business Cycle." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 4562, December 1993.

Chatterjee, Satyajit; Cooper, Russell W. and Ravikumar, B. "Strategic Complementarity in Business Formation: Aggregate Fluctuations and Sunspot Equilibria." *Review of Economic Studies,* October 1993, *60*(4), pp. 795–811.

Cooper, Russell W. and John, Andrew. "Coordinating Coordination Failures in Keynesian Models." *Quarterly Journal of Economics,* August 1988, *103*(3), pp. 441–63.

Diamond, Peter. "Aggregate Demand Management in Search Equilibrium." *Journal of Political Economy,* October 1982, *90*(5), pp. 881–94.

Drugeon, Jean Pierre and Wigniolle, Bertrand. "Continuous-Time Sunspot Equilibria and Dynamics in a Model of Growth." *Journal of Economic Theory,* April 1996, *69*(1), pp. 24–52.

Evans, George W. and Honkapohja, Seppo. "On the Local Stability of Sunspot Equilibria under Adaptive Learning Rules." *Journal of Economic Theory,* October 1994, *64*(1), pp. 142–61.

———. "Adaptive Learning and Expectational Stability: An Introduction," in A

---

[9] In fact, the algorithm needs to be supplemented with a "projection facility" to obtain local convergence with probability one. Evans and Honkapohja (1998) show that without a projection facility one can still obtain positive convergence results.

Kirman and M. Salmon, eds., *Learning and rationality in economics.* Oxford: Blackwell, 1995a, pp. 104–26.

———. "Local Convergence of Recursive Learning to Steady States and Cycles in Stochastic Nonlinear Models." *Econometrica,* January 1995b, *63*(1), pp. 195–206.

———. "Economic Dynamics with Learning: New Stability Results." *Review of Economic Studies,* January 1998, *65*(1), pp. 23–44.

Farmer, Roger. "Sticky Prices." *Economic Journal,* November 1991, *101*(409), pp. 1369–79.

Farmer, Roger and Guo, Jang Ting. "Real Business Cycles and the Animal Spirits Hypothesis." *Journal of Economic Theory,* June 1994, *63*(1), pp. 42–72.

Gali, Jordi. "Monopolistic Competition, Business Cycles and the Composition of Aggregate Demand." *Journal of Economic Theory,* June 1994, *63*(1), pp. 73–96.

———. "Product Diversity, Endogenous Markups, and Development Traps." *Journal of Monetary Economics,* December 1995, *36*(1), pp. 39–63.

Guesnerie, Roger and Woodford, Michael. "Endogenous Fluctuations," in Jean-Jacques Laffont, ed., *Advances in economic theory Sixth World Congress,* Vol. II. Cambridge: Cambridge University Press, 1992, pp. 289–412.

Hamilton, James D. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica,* March 1989, *57*(2), pp. 357–84.

Howitt, Peter and McAfee, R. Preston. "Animal Spirits." *American Economic Review,* June 1992, *82*(3), pp. 493–507.

Kehoe, Timothy J.; Levine, David K. and Romer, Paul M. "On Characterizing Equilibria of Economies with Externalities and Taxes as Solutions to Optimization Problems." *Economic Theory,* January 1992, *2*(1), pp. 43–68.

Lucas, Robert E., Jr. "Adaptive Behavior and Economic Theory." *Journal of Business,* October 1986, Pt. 2, *59*(4), pp. S401–26.

Matsuyama, Kiminori. "Complementarities and Cumulative Processes in Models of Monopolistic Competition." *Journal of Economic Literature,* June 1995, *33*(2), pp. 701–29.

Romer, Paul M. "Growth Based on Increasing Returns Due to Specialization." *American Economic Review,* May 1987 (*Papers and Proceedings*), *77*(2), pp. 56–62.

Schmitt-Grohé, Stephanie. "Comparing Four Models of Aggregate Fluctuations Due to Self-Fulfilling Prophecies." *Journal of Economic Theory,* January 1997, *72*(1), pp. 96–147.

Shleifer, Andrei. "Implementation Cycles." *Journal of Political Economy,* December 1986, *94*(6), pp. 1163–90.

Silvestre, Joaquim. "The Market-Power Foundations of Macroeconomic Policy." *Journal of Economic Literature,* March 1993, *31*(1), pp. 105–41.

Startz, Richard. "Growth States and Sectoral Shocks." Mimeo, University of Washington, January 1994.

Woodford, Michael. "Self-Fulfilling Expectations and Fluctuations in Aggregate Demand," in N. Gregory Mankiw and David Romer, eds., *New Keynesian economics,* Vol. 2. Cambridge, MA: MIT Press, 1991, pp. 77–110.

Xie, Danyang. "Divergence in Economic Performance: Transitional Dynamics with Multiple Equilibria." *Journal of Economic Theory,* June 1994, *63*(1), pp. 97–112.

Young, Alwyn. "Substitution and Complementarity in Endogenous Innovation." *Quarterly Journal of Economics,* August 1993, *108*(3), pp. 775–808.

Financial Intermediation and Regime Switching in Business Cycles

By COSTAS AZARIADIS AND BRUCE SMITH*

We study a one-sector growth model where capital investment is credit financed, and there is an adverse selection problem in credit markets. The presence of adverse selection creates an indeterminacy of equilibrium. Many equilibria display permanent fluctuations characterized by transitions between Walrasian regimes and regimes of credit rationing. Cyclical contractions involve declines in real interest rates, increases in credit rationing, and withdrawals of savings from banks. For some configurations of parameters all *equilibria display cyclical fluctuations. We provide sufficient conditions for deterministic cycles consisting of m periods of expansion followed by n periods of contraction to exist.* (*JEL* E30, E32)

Monetary economists have frequently expressed the view that the financial system is an important source of—and propagation mechanism for—cyclical fluctuations. Indeed, John Maynard Keynes (1936), Henry Simons (1948), Milton Friedman (1960), and many others have argued that the free and unregulated operation of financial markets can lead to indeterminacy of equilibrium and "excessive economic fluctuations," even in the absence of shocks impinging on the rest of the economy. In modern terms, this argument claims that the financial system itself is a source of endogenously arising economic volatility.

This view has a strong empirical foundation. Most of the pre-World War II recessions were associated with substantial transfers of resources out of the banking system and into other assets. For instance, almost all of the pre-World War II recessions described by Friedman and Anna J. Schwartz (1963) were associated with increases in the currency-deposit ratio, and with an implied withdrawal of resources from the banking system. Particularly severe recessions were associated with particularly sharp increases in the currency-deposit ratio (that is, with bank panics). And even in the last three decades, several recessions have been accompanied by phenomena termed "disintermediation" or "credit crunches." In all of these episodes the volume of bank-extended credit declined, and "credit crunches" have often been associated with an increased incidence of nonprice rationing of credit.[1]

Why do we observe such sharp fluctuations in the volume of intermediated credit? Can these fluctuations be the "cause" of business cycles, as many (for instance, Friedman and Schwartz) have claimed, or are they merely

---

* Azariadis: Department of Economics, University of California, Los Angeles, CA 90024; Smith: Department of Economics, University of Texas, Austin, TX 78712. A preliminary version of this paper was entitled "Endogenous Intermediation Cycles." For helpful comments we are indebted, without implication, to Patrick Asea, Angel de la Fuente, Jean-Michel Grandmont, Nobuhiro Kiyotaki, Simon Potter, Pietro Reichlin, Bill Schworm, two anonymous referees, and to seminar audiences at the University of Southern California, the NBER Summer Institute, the Federal Reserve Banks of St. Louis and Minneapolis, the University of California at Davis, Irvine, Riverside, and San Diego, the University of British Columbia, Simon Fraser University, the University of Victoria, the Canadian Macroeconomic Theory Meeting, Stanford University, and Stanford Institute for Theoretical Economics (SITE). Azariadis acknowledges hospitality from the Instituto de Analisis Economico, Universitat Autonoma de Barcelona, Universita di Napoli, Universidad Carlos III, and travel support from the Human Capital and Mobility Program, Commission of the European Communities.

[1] See, for example, Stacey L. Schreft (1990) or Schreft and Raymond F. Owens (1995).

symptom of some more general phenomenon? If credit market activity does contribute to economic fluctuations, where does the impulse arise, and how is it transmitted to the real sector? And if the financial system is itself a source of endogenous shocks, are fluctuations in real economic activity merely a *possible* consequence or are they completely inevitable? For instance, can it be the case that under some circumstances the financial system *must* suffer from endogenous volatility which then spreads to the rest of the economy?

To answer these questions, this paper considers the relationship between credit and production in a simple model of dynamic general equilibrium, namely the nonmonetary overlapping generations economy with production introduced by Peter A. Diamond (1965).[2] We modify that economy in only three respects: we introduce some intragenerational heterogeneity, we force some kinds of capital investment to be credit financed, and we allow for the existence of an adverse selection problem in capital markets. Under complete public information about the characteristics of potential borrowers, we show that the first two of these modifications are purely cosmetic, and make no qualitative difference to the properties of competitive equilibria. Specifically, given any positive initial capital stock, the economy monotonically approaches a nontrivial steady-state equilibrium, which is unique under our assumptions. Thus equilibrium is determinate, and endogenous fluctuations cannot arise.

Heterogeneity, however, makes a lot of difference when there is private information about borrower characteristics (here *ex ante* information about loan repayment probabilities). In the presence of this adverse selection problem, lenders will seek to elicit from borrowers information regarding their loan repayment probabilities. Lenders will do so by structuring the loan contracts they offer—which specify both loan quantities and interest rates—to induce potential borrowers to self-select or, in effect, to reveal their type. Thus all loan contracts offered must be incentive compatible.

In our economy there is a range of values for the current capital stock (equivalently, for factor prices) which make the full-information allocation of credit incentive compatible. Here Walrasian allocations are competitive equilibria, and the equilibrium law of motion of the capital stock coincides with the law of motion that would prevail under perfect information. However, we also show that there is a range of current capital stocks for which Walrasian allocations cannot be incentive compatible; here incentive constraints must be binding and credit will be rationed. Credit rationing in our economy necessarily diminishes capital formation and reduces the level of real activity relative to what would be observed under public information. This raises two distinct possibilities. One is that the allocation of credit that would arise under public information is incentive compatible, or consistent with self-selection, under private information. In this case the full-information allocation can be duplicated by the appropriate choice of loan contracts, despite the presence of the informational asymmetry. As a result, a fully Walrasian allocation is feasible, and the adverse selection problem is innocuous. Alternatively, it may be impossible to induce self-selection without imposing quantity constraints on contracts. In this case the adverse selection problem "matters": lenders will use credit rationing as a means of separating borrowers with different *ex ante* repayment probabilities, and incentive constraints will be binding in equilibrium.

Finally, there is a nontrivial closed interval of current capital stocks having the following property: if the public-information allocation arises, it is incentive compatible and hence a true competitive equilibrium. At the same time, if credit rationing occurs, incentive constraints do bind on the choice of equilibrium loan contracts, so that a non-Walrasian equilibrium also exists in which the adverse selection problem does affect allocations. For this range of current capital stocks, then, the economy can be in either of two equilibrium regimes: a Walrasian regime in which competitive markets allocate credit in the standard way, or a regime of credit rationing. When credit is rationed, resources leave the banking system and the allocation of investment becomes less efficient in a manner we will make precise. Since each of these regimes is consistent with equilibrium, equilibrium is indeterminate: the economy can follow either the full-information or the private-information law of motion for the capital stock. Moreover, as we will demonstrate, there exist equilibria in which the economy can switch from one law of motion to the other in either a deterministic or a stochastic manner. These regime transitions will be associated with fluctuations in output and the capital stock which need not dampen over time. Thus both indeterminacy of equilibrium and "excessive" fluctuations can be observed when agents are privately informed about loan repayment probabilities.

---

[2] Azariadis and Smith (1996a) study the role of fiat money or national debt in a neoclassical growth model with adverse selection.

Why does private information induce this regime switching, and the economic fluctuations that go with it? As we will show, Walrasian allocations are consistent with an incentive-compatible allocation of credit *only if the real rate of return on savings (deposits) is sufficiently high.* Then suppose the economy is currently in the Walrasian regime. If savers (depositors) anticipate a sufficiently low real return, they transfer some savings out of the banking system and into other, lower-yielding assets. This savings outflow forces banks to ration credit, firms to curtail capital investment, and real economic activity to decline. Moreover, the existence of credit rationing breaks the link between the marginal product of capital and the equilibrium rate of interest. This makes it possible for interest rates to fall in the way that savers anticipate. Thus depositors' expectations of declining interest rates become a self-fulfilling prophecy which forces a transition from the Walrasian regime into a regime of credit rationing. The consequence of tighter credit is an economic contraction in which both real interest rates and real activity decline.

Alternatively, suppose that the economy is currently in a regime of credit rationing. As we will demonstrate, *credit rationing can occur only if the real rate of return on savings (deposits) is sufficiently low.* If depositors then expect rising real interest rates, resources will flow back into the banking system. Credit can no longer be rationed, banks will finance more capital investment, and real activity will expand. Here changes in depositors' expectations act once more as a self-fulfilling prophecy of rising real interest rates, rising real activity, and a transition from a regime of credit rationing to one of Walrasian allocations.

As this discussion suggests, depositors' beliefs about financial regimes are at the heart of the multiplicity of equilibrium and of the regime transitions that can be observed here. This observation leads us to ask an additional question: are there any equilibria in which depositor beliefs never change? We describe some conditions under which the answer is yes. Under these conditions, there is necessarily a large set of equilibria. In two of these equilibria, depositor beliefs are invariant over time, and depositors always expect high (low) real interest rates. As a consequence, the economy is always in a Walrasian (credit-rationed) equilibrium. In addition, there is a large set of equilibria—which we characterize—where depositors' beliefs fluctuate in ways that induce regime transitions. Here it is possible to take the view that depositor beliefs fluctuate somewhat capriciously.

However, we also state conditions under which this turns out to be an overstatement, and there are *no* equilibria consistent with unchanging depositor expectations. When these conditions are satisfied, *all* equilibria display regime transitions and endogenously arising volatility.

Why are regime transitions a necessary feature of any equilibrium for some economies? As we have already argued, there is a lower bound on the rate of interest that is consistent with a Walrasian allocation. Moreover, in a Walrasian allocation, the real rate of interest and the marginal product of capital coincide. Thus, if the capital stock is too large (the marginal product of capital is too small) in a Walrasian allocation, this allocation cannot constitute an equilibrium for it is not consistent with an incentive-compatible allocation of credit. There is, then, an upper bound on the level of real activity that can be observed in the Walrasian regime.

Similarly, we show that there is an upper bound on the marginal product of capital (a lower bound on the level of real activity) that is consistent with the rationing of credit. If the capital stock is too low, banks will find bigger profits in lending indiscriminately to all borrowers at a common interest rate rather than separating them into "good" and "bad" risks. Therefore, rationing cannot occur if the capital stock is too low. This places a floor on the level of real economic activity.

Suppose that an economy is always in the Walrasian (credit-rationed) regime. Under our assumptions, the capital stock will monotonically approach its public-information (private-information) steady-state level. If this exceeds (is below) the upper (lower) bound on real activity that is consistent with a Walrasian (rationed) allocation of resources, then the implied time path of the economy is not an equilibrium outcome. Regime transitions must occur, and depositors must periodically revise their beliefs. Of course as this discussion suggests, if the full-information (private-information) steady-state capital stock is sufficiently low (high), there will be equilibrium paths where the same financial regime can prevail indefinitely. Here regime transitions may, but need not, occur.

For each of these cases, we examine the existence of deterministic perfect-foresight equilibria in which there are m periods of expansion (in a Walrasian regime), followed by n periods of contraction (in a regime of credit rationing). When it is possible to remain in each regime indefinitely, we show that such an equilibrium exists for *every* pair of integers (m, n). Moreover, all such equilibria are asymptotically stable, so that the economy generates a high-dimensional indeterminacy of equilibrium. In addition, business cycles are potentially asymmetric—in the sense that expansions are longer (shorter) than contractions—if $m > (<)\ n$.

When it is *not* possible for either regime to prevail indefinitely, we are not guaranteed that cyclical equilibria exist for arbitrary (m, n) pairs. In this case we describe conditions under which a deterministic equilibrium (m, n) cycle does exist, for some pairs (m, n). We also focus our attention on the following question: is there an equilibrium cycle along which depositor beliefs are revised as infrequently as possible? We call such equilibria *maximally persistent* (m, n) cycles, and we state sufficient conditions for maximally persistent (m, n) cycles to exist. We also show how to find the associated values of m and n, and we indicate when expansions will be longer or shorter than contractions in a maximally persistent cycle.

The remainder of the paper proceeds as follows. Section I lays out the environment and the nature of trades, and describes the equilibrium conditions that obtain when credit is and is not rationed. Section II shows how the economy can transit between Walrasian regimes and regimes of credit rationing, while Sections III and IV examine the existence of cyclical equilibria which display undamped oscillations. Section V concludes.

I. Capital Accumulation With Adverse Selection

A. *Environment*

We consider a simple variant of Diamond's one-sector neoclassical growth model. In particular, at each date $t = 0, 1, \cdots$ a new set of two-period lived, overlapping generations is born. Each generation is identical in composition, and consists of a continuum of agents of measure one.

At each date there is a single consumption good, which is produced using a standard constant-returns-to-scale production function with capital and labor as inputs. A capital input of K_t combined with a labor input of L_t permits $F(K_t, L_t)$ units of this good to be produced at t. We let $k_t \equiv K_t/L_t$ denote the capital-labor ratio, and $f(k_t) \equiv F(k_t, 1)$ denote the intensive production function. We assume that $f(0) = 0$, that $f'(k) > 0 > f''(k)\ \forall k \geq 0$, and that f satisfies the usual Inada conditions. We also assume that the consumption good can be stored: one unit of the good stored at t returns a units of the good at $t + 1$. Throughout we think of a as being relatively small, so that storage is a relatively unproductive activity. Finally, the consumption good can be used to produce capital; one unit of consumption placed in a capital investment at t returns one unit of capital at $t + 1$.

Within each generation, agents are divided into two types. A fraction $\gamma > 0.5$ of young agents are of type 1. These agents are endowed with one unit of labor when young, and are retired when old. In addition, they have access to the storage technology just described, but they have *no* access to the technology for converting current goods into future capital. A fraction $1 - \gamma < 0.5$ of each generation is of type 2. Type 2 agents cannot work when young, but are endowed with one unit of labor when old. In addition, we assume that type 2 agents have no access to the goods storage technology, but that they are endowed with the technology for converting current consumption into future capital. Thus, in many respects, type 1 and 2 agents are mirror images of one another. Finally, we assume that all agents care only about old-age consumption, and are risk neutral.[3] In particular, labor generates no disutility.

Notice that type 1 agents are natural lenders at any interest rate because they need to provide for old-age consumption. Type 2 agents are natural borrowers who need credit to finance capital investments. Between borrowers and lenders stand financial intermediaries, or banks, which accept deposits and extend loans. Their cost of doing so is zero.[4]

Finally, there is a set of initial old agents who are endowed with a per capita capital stock of $k_0 > 0$. We assume that capital depreciates completely in production.

B. *Full Information*

In this subsection we describe a competitive equilibrium of this economy under the assumption of full information, and in particular that all allocations and the type of each agent are publicly observable. We also assume throughout, without loss of generality, that each type 2 agent (recall that these agents own capital) runs one firm and works for himself.

At date t each young type 1 agent supplies one unit of labor inelastically, earning the real wage rate w_t. All of this income is saved, to be allocated between storage and bank deposits. We let s_t be per capita storage at t, so $\gamma w_t - s_t$ is per capita deposits. Deposits earn the competitive gross return R_{t+1} between t and $t + 1$, which both banks and depositors treat as parametric.

Each young type 2 agent (firm) borrows b_t and uses it to produce a per firm capital stock of K_{t+1} at t. Loan market clearing, then, requires that

$$(1) \qquad (1 - \gamma)b_t + s_t = \gamma w_t.$$

In addition, there will be a positive supply of deposits if and only if (iff) these weakly dominate storage in rate of return, so that

$$(2) \qquad a \leq R_t, \quad \forall t.$$

Old type 2 agents at t have the inherited (per firm) capital stock K_t, which they combine with their own labor, plus N_t units of young labor in order to produce output. In addition these agents have an inherited interest obligation of $r_t b_{t-1} = r_t K_t$, where r_t is the gross loan rate of interest, between $t - 1$ and t, charged by banks. Then old type 2 agents have an income (consumption) level of

$$F(K_t, 1 + N_t) - w_t N_t - r_t K_t,$$

which they maximize with respect to N_t. Hence if total labor input for the firm is $L_t \equiv 1 + N_t$, we have

$$(3) \quad w_t = F_2(K_t, L_t)$$

$$\equiv f(k_t) - k_t f'(k_t) \equiv w(k_t); \quad t \geq 0$$

Note that $w'(k) > 0$ holds for all k, and in addition we will assume that.

ASSUMPTION 1:

$$w''(k) < 0.$$

The above assumption holds if, for instance, F is any constant-elasticity-of-substitution function with elasticity of substitution no less than one.

We assume that there is free entry into banking, so that

$$(4) \qquad r_t = R_t$$

---

[3] The assumption that agents are risk neutral precludes lotteries from playing a useful role under private information. The assumption that agents care only about old-age consumption allows us to abstract from consumption/savings decisions. Both are inessential simplifications.

[4] It is not essential that credit extension be intermediated. However, the assumption that all credit is intermediated entails no loss of generality.

holds for all $t \geq 0$. In addition, competition among banks for borrowers implies that banks must offer type 2 agents at t the loan quantity which maximizes their lifetime utility: $F(b_t, 1 + N_{t+1}) - w_{t+1}N_{t+1} - R_{t+1}b_t$. Hence

$$R_{t+1} = F_1(b_t, L_{t+1}) = F_1(K_{t+1}, L_{t+1}) = f'(k_{t+1}); \quad t \geq 0. \tag{5}$$

Finally, since a fraction $1 - \gamma$ of the population is firms and γ is workers at each date, clearly $N_t = \gamma/(1 - \gamma)$, and $L_t = 1 + N_t = 1/(1 - \gamma)$. Thus

$$k_t = K_t / L_t = (1 - \gamma)K_t = (1 - \gamma)b_{t-1}; \quad t \geq 1. \tag{6}$$

It is now straightforward to describe the equilibrium law of motion for the capital stock. Equations (1), (3), and (6) imply that

$$k_{t+1} = \gamma w(k_t) - s_t. \tag{7}$$

Suppose that $f'(k_{t+1}) > a$. Then $s_t = 0$ holds, and (7) becomes

$$k_{t+1} = \gamma w(k_t). \tag{7'}$$

Under our assumptions, (7′) describes an increasing, concave locus that passes through the origin, as depicted in Figure 1. If $\gamma w'(0) > 1$ holds, then there is a unique nontrivial steady-state equilibrium, denoted k^0 in Figure 1, Panel (a).[5] Given the initial per capita capital stock k_0, there is a unique sequence $\{k_t\}$ that monotonically converges to the steady state. Thus, under full information, this economy can display neither indeterminacy nor economic fluctuations.

C. Credit Rationing

Next we introduce private information and examine what occurs when incentive constraints bind. Then, in Section II, we examine a full general equilibrium in which incentive constraints may or may not bind in an endogenous way.

We work with a simple structure of information; household type and storage activity are private information, while all market (both labor and credit market) transactions are publicly observed. These assumptions imply that young type 2 agents cannot credibly claim to be type 1 when young, since they are unable to supply labor. Type 1 agents, however, can borrow when young. If they do so, they must borrow the same amount (b_t) as type 2 agents, and they cannot work. In addition, type 1 agents are unable to produce capital, and hence cannot function as producers in old age. Thus any type 1 agent who borrows will be discovered as having misrepresented his type. In order to avoid punishment, a dissembling type 1 agent will simply store what he borrows and "go underground" or abscond with his loan. Thus a loan made to a type 1 agent will never be repaid.[6]

Given these circumstances, bank behavior must be modified in two respects. First, loan contracts must earn nonnegative profits. Let μ_t denote the fraction of type 1 agents who claim to be of type 2. Since such agents never repay their loans (and all borrowers borrow the same amount), banks earn nonnegative profits iff

$$r_{t+1} \geq R_{t+1}\{1 + \mu_t[\lambda/(1 - \lambda)]\}; \quad t \geq 0. \tag{8}$$

Equation (8) requires that type 2 agents, who actually repay loans, pay enough to cover the

---

[5] [illegible], of course, assumes that $f'(k^0) > a$.

[6] The hallmark of any model of credit rationing based on adverse selection or moral hazard [for example, Joseph Stiglitz and Andrew Weiss (1981) or Valerie R. Bencivenga and Smith (1993)] is that some borrowers have a higher probability of repayment than others, and hence care more about the interest rate dimension of the loan contract. Our specification is simply the most extreme (and simplest) version of this possibility: type 1 agents default on loans with probability one, and type 2 agents repay with the same probability. There is no conceptual difficulty associated with allowing type 1 and 2 agents to have loan repayment probabilities strictly between zero and one. However, this merely adds complexity without bringing any additional substantive issues into the analysis.

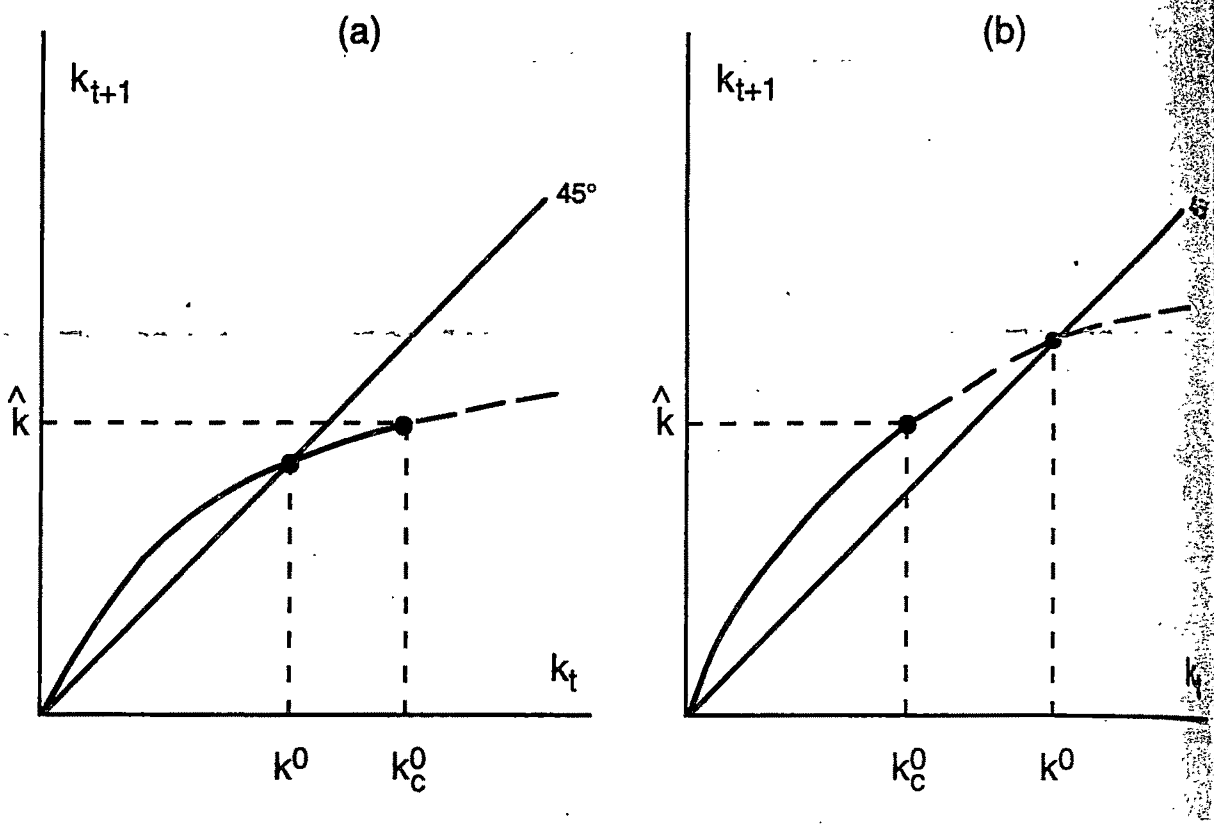


FIGURE 1. WALRASIAN EQUILIBRIUM LAW OF MOTION FOR k_t

defaults associated with loans to type 1 agents. Second, if $\mu_t < 1$ holds,[7] type 1 agents must do at least as well by working when young and saving as they would by claiming to be of type 2. Working when young and saving generates a lifetime utility level of $R_{t+1}w_t$ at t, while claiming to be of type 2 generates a lifetime utility level of ab_t. Hence loan contracts must satisfy the incentive constraint (or self-selection condition)[8]

$$(9) \qquad R_{t+1}w_t \geq ab_t; \qquad t \geq 0.$$

Since equation (9) is central to the analysis that follows, it is important to be clear about its content, and about the assumptions that support it. Equation (9) simply states the condition that is required in order to prevent type 1 agents from falsely misrepresenting their type, borrowing, and absconding with their loans. The essential assumptions underlying it are: (i) agents' types are private information; (ii) the amount borrowed is publicly observed, while (iii) the allocation of funds between investment and storage is not; and (iv) the returns on investment cannot be concealed, while the returns on storage can.

Assumption (i), of course, simply asserts that lenders confront an adverse selection problem, while assumption (ii) is quite standard in models of adverse selection. Assumption (iii) allows borrowers to divert funds from their "intended" uses; this is a common formulation in models of investment under moral hazard.[9] Assumption (iv) is based on the notion that some activities—like formal production utilizing a substantial capital input—are of a "public" nature and their outcome is difficult to conceal, while others are of a "private" nature which makes it easier to misappropriate their returns. Similar assumptions on return appropriability are common in models of moral hazard (Gertler and Hubbard, 1988; Gertler and Rogoff, 1990).

---

[7] If $\mu_t = 1$, no young agents work at t, no savings are supplied, and $k_{t+1} = 0$. The economy jumps to the autarkic steady state.

[8] Our assumption that labor-market activity is perfectly observable implies that dissembling type 1 agents can supply no labor when young. However, there is no conceptual difficulty with allowing dissembling type 1 agents to generate some labor income when young. In particular, suppose that a young type 1 agent who misrepresents his type can supply $\phi < 1$ units of labor, where $1 - \phi$ then represents the cost of working surreptitiously. Then equation (9) becomes $R_{t+1}w_t \geq a(b_t + \phi w_t)$, while equation (12) below must be replaced by $k_{t+1} = (1 - \gamma)(1 - \phi)w(k_t)$. Clearly this leads to no qualitative difference in results. We therefore pursue the simplest specification with $\phi = 0$.

Following Michael Rothschild and Joseph Stiglitz (1976), we assume that intermediaries are Nash competitors in loan markets who take the deposit rate R_t and the announced loan contracts of other banks as given. As in Rothschild and Stiglitz, it is easy to show that any Nash equilibrium contract earns zero profits, so that (8) holds with equality. In addition, it is possible to show that any nontrivial equilibrium (that is, any equilibrium with $k_t > 0\ \forall t$) has $\mu_t = 0\ \forall t$. In particular, contracts induce self-selection, and pooling is not a possibility.[10]

Thus, if an equilibrium with credit rationing exists, it continues to satisfy equations (3) and (6). Therefore (9) reduces to

$$R_{t+1}w(k_t) \geq ak_{t+1}/(1-\gamma). \tag{9'}$$

In addition, it is necessary that type 2 agents be willing to borrow; it is easy to show that they are if and only if (iff) the marginal product of capital exceeds the loan rate. In short, the inequality

$$f'(k_{t+1}) \geq R_{t+1} \tag{10}$$

must hold. If (10) holds as a strict inequality, borrowers would like to borrow arbitrarily large amounts; in this case the incentive constraint (9) [or (9′)] holds as an equality and determines the equilibrium loan quantity. Thus, when credit rationing obtains, (9′) holds as an equality.

Finally, in order for announced loan contracts (b_t, r_{t+1}) to constitute a Nash equilibrium, it must be the case that no intermediary can offer an alternative contract $(\tilde{b}_t, \tilde{r}_{t+1})$ which is preferred by type 2 agents, and which satisfies (8) for some μ_t. It is straightforward to show that, if there is such a contract, it must be a pooling contract ($\mu_t = 1$) with $\tilde{r}_{t+1} \geq R_{t+1}/(1-\gamma)$ in order to satisfy (8). Moreover, a pooling contract will attract type 2 agents iff $R_{t+1}/(1-\gamma) \geq f'(k_{t+1})$ holds.[11] Since it is impossible to observe a pooling contract in a nontrivial equilibrium, the existence of an equilibrium with credit rationing requires that there be no pooling contract which type 2 agents prefer to the contract $(b_t, R_{t+1}) = (R_{t+1}w_t/a, R_{t+1})$. This is so if

$$f'(k_{t+1}) \leq R_{t+1}/(1-\gamma). \tag{11}$$

To summarize: in a regime of credit rationing, equations (3), (4), and (6) hold, as does (9′) at equality. Equations (10) and (11) must hold as well; these may (and typically will) be inequalities.

Prior to proceeding, it will be useful to have the following preliminary result. Its proof appears in Azariadis and Smith (1996b).

PROPOSITION 1: *Suppose that credit rationing occurs at t, and that $R_{t+1} > a$. Then*

$$k_{t+1} = \gamma w(k_t).$$

It is an immediate corollary of Proposition 1 that credit rationing has no bearing on the aggregate behavior of the economy unless $R_{t+1} = a$. We therefore focus our attention in the remainder of this section on equilibria in which there is credit rationing, and in which $R_{t+1} = a$ (real interest rates are low). The following proposition states our main result.

---

[9] See, for instance, Mark Gertler and R. Glenn Hubbard (1988), or Mark Gertler and Kenneth Rogoff (1990). Parenthetically, it is not essential to our results that type 1 agents are completely incapable of producing capital, or that type 2 agents are completely incapable of storing goods. All that is really required is that each type of agent is so inefficient in the appropriate activity that they are never tempted to undertake it.

[10] See Azariadis and Smith (1996a) for a proof in this particular context. Parenthetically, the argument they give depends on banks treating the deposit rate R_t parametrically. See David Kreps (1990 Ch. 17) for the strategic foundations of this formulation.

[11] Again, see Azariadis and Smith (1996a) for a formal proof.

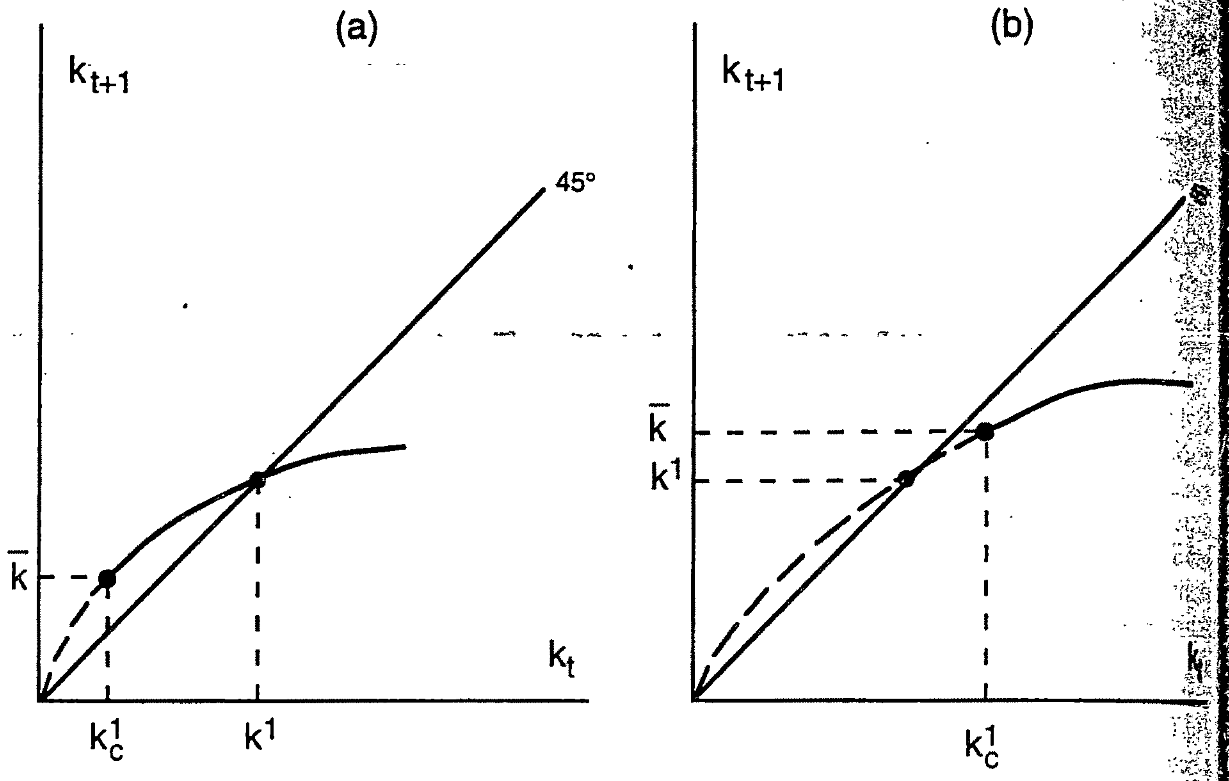


FIGURE 2. CREDIT RATIONING
LAW OF MOTION FOR k_t

PROPOSITION 2: *Suppose that there is credit rationing at t, and that* $R_{t+1} = a$. *Then the evolution of the capital stock is described by*

$$k_{t+1} = (1 - \gamma)w(k_t), \tag{12}$$

while the aggregate (per capita) quantity of storage at t satisfies

$$s_t = (2\gamma - 1)w(k_t) > 0. \tag{13}$$

Proposition 2 follows from substituting $R_{t+1} = a$ into (9′) to obtain (12), and from substituting (12) into (7), to obtain (13). The proposition asserts that, when credit rationing is accompanied by low real returns on savings, savings flow out of the banking system into the unintermediated activity we call "storage."

As noted in Proposition 2, when credit rationing does exist at date t, and when $R_{t+1} = a$ holds, equation (12) governs the evolution of the capital stock. Equation (12), of course, defines an increasing, concave function (as depicted in Figure 2) which has a unique nontrivial intersection with the 45-degree line if $(1 - \gamma)w'(0) > 1$. We denote this intersection by k^1. Since $\gamma > 0.5$, the law of motion defined by (12) lies everywhere below the law of motion defined by (7)—so that credit rationing impedes capital formation—and $k^0 > k^1$ necessarily holds. Finally we note that when credit is rationed, and when $R_{t+1} = a$, inequality (11) reduces to

$$f'(k_{t+1}) \leq a/(1 - \gamma). \tag{14}$$

Thus (14) must hold at any t for which there is credit rationing.

If credit is rationed at every date, then (12) gives the equilibrium sequence of capital stocks starting from k_0. Clearly this sequence will monotonically approach k^1, so that all equilibria are unique and display monotonic dynamics. Multiple equilibria and endogenous fluctuations are associated with transitions between Walrasian regimes and regimes with credit rationing. Such transitions are the focus of the next section.

II. Endogenous Regime-Switching Mechanisms

Suppose that, given the inherited capital-labor ratio k_t at t, the value of k_{t+1} yielded by (7′) satisfies $f'(k_{t+1}) > a$, and in addition that (k_t, k_{t+1}) satisfies (9′). Then the full-information allocation is incentive compatible, and constitutes an equilibrium at t. Alternatively, suppose that, given k_t, equation (9′) gives a value k_{t+1} satisfying both $f'(k_{t+1}) \geq a$ and (14). Then credit rationing is consistent with an equilibrium outcome at t in which $R_{t+1} = a$.

These two possibilities define two distinct regimes for each time period; in regime 0, the full-information allocation is incentive compatible and a Walrasian equilibrium obtains; in regime 1 the incentive constraint binds, R_{t+1} equals a, and credit to borrowers is rationed. A brief description of each regime follows.

Regime 0—In order to describe when the economy can be in regime 0, it is useful to define two critical values for the capital stock. Let $\hat{k}$ satisfy

$$f'(\hat{k}) \equiv a\gamma/(1-\gamma), \tag{15a}$$

while k_c^0 is defined by

$$\hat{k} \equiv \gamma w(k_c^0). \tag{15b}$$

Then $\hat{k}$ is the largest "full-information capital stock" that can satisfy the incentive constraint, and k_c^0 is the capital stock that maps into $\hat{k}$ under (7′). These values are crucial ingredients in the analysis that follows; we now briefly elaborate on their significance.

Equations (7′) and (9′) indicate that the incentive constraint must be binding whenever the real rate of interest is too low. Thus large values of the capital stock are inconsistent with a Walrasian allocation because they imply excessively low interest rates. At these interest rates it is not possible to deter type 1 agents from misrepresenting their type without the existence of some credit rationing. Thus $k_{t+1}(k_t)$ cannot exceed $\hat{k}(k_c^0)$ without inducing some rationing of credit at time t.

Given the values of $(\hat{k}, k_c^0)$, it is easy to check that any sequence (k_t, s_t, R_t) conforming to the initial condition k_0 is an equilibrium without credit rationing if, for each t,

$$k_{t+1} = \gamma w(k_t); \tag{16a}$$

$$s_t = 0; \tag{16b}$$

$$R_t = f'(k_t) > a; \tag{16c}$$

$$k_{t+1} < \hat{k} \text{ or } k_t < k_c^0. \tag{16d}$$

Equation (16) defines equilibrium sequences for which borrower characteristics are effectively public information. Solution sequences of (16a) converging to k^0 will be competitive equilibria for regime 0 if they are bounded above by the critical value k_0^c defined in equations (15a) and (15b) and shown in Figure 1.

We conclude that it is feasible to be in a Walrasian regime if two conditions are met: (i) the level of economic activity is not too high (relative to k_c^0); and (ii) the investment projects available to low-quality borrowers (type 1 agents) are sufficiently less productive than those open to high-quality borrowers (type 2 agents).

Regime 1—Again, in order to describe when the economy can be in regime 1, it is necessary to define two more critical values of the capital stock. Let $\bar{k}$ and k_c^1 satisfy:

$$f'(\bar{k}) = a/(1-\gamma); \tag{17a}$$

$$\bar{k} = (1-\gamma)w(k_c^1), \tag{17b}$$

respectively. Then $\bar{k}$ is the smallest capital stock consistent with the existence of a separating Nash equilibrium under private information [see equation (14)], and k_c^1 is the capital stock that maps into $\bar{k}$ under (9′). From (15a) and (17a) we obtain

$$\bar{k} < \hat{k}. \tag{18}$$

As before, the values $\bar{k}$ and k_c^1 play an essential role in the subsequent analysis, so we again comment on their significance. Equation (14) indicates that, if the marginal product of capital is too high relative to the rate of interest (and $R_{t+1} = a$ holds in regime 1), it is not possible to deter lenders from pooling type 1 and type 2 agents, and lending to them indiscriminately. This outcome is inconsistent with the existence of a nontrivial equilibrium; hence the existence of an equilibrium within regime 1 requires that $k_{t+1} \geq \bar{k}$ ($k_t \geq k_c^1$) hold.

Given the values ($\bar{k}$, k_c^1), any sequence (k_t, s_t, R_t) that starts from a given initial condition k_0 is a competitive equilibrium with credit rationing if, for each t, it satisfies the following conditions:

$$k_{t+1} = (1 - \gamma)w(k_t); \tag{19a}$$

$$s_t = (2\gamma - 1)w(k_t); \tag{19b}$$

$$R_{t+1} = a < f'(k_{t+1}); \tag{19c}$$

$$k_{t+1} > \bar{k} \text{ or } k_t > k_c^1. \tag{19d}$$

As previously, the locus defined by (19a) in Figure 2 lies entirely below the locus defined by (16a), and its fixed point, k^1, lies below k^0.

Figure 2 also depicts the domain of definition, $[k_c^1, \infty)$, for equation (19a) and depicts two possible cases: the unique positive fixed point, k^1, of this equation is in that domain when Panel (a) obtains, but not when Panel (b) does. Given any initial value k_0 of the capital stock, separating competitive equilibria with credit rationing and $R_{t+1} = a$ are more likely to exist if the critical value k_c^1 is low relative to the fixed point k^1, i.e., if: (i) the level of economic activity is not too low (relative to k_c^1); and (ii) the investment projects available to low-quality borrowers are not much less productive than the ones open to high-quality borrowers.

A. *Regime Transitions*

The previous section considered equilibria in which either regime 0 or regime 1 prevails permanently. However, it is also possible—or even necessary—that there exist equilibria in which the economy transits between these two regimes in either a deterministic, or a stochastic manner. This subsection discusses somewhat informally how and why an economy may shift from a Walrasian regime of slack incentive constraints to one of credit rationing and tight incentive constraints. Sections III and IV treat the same issues more formally and in greater depth. We start by amalgamating Figures 1 and 2 into the top two panels of Figure 3. In Figure 3(a) the fixed point of each regime is contained in its domain of definition; in other words, Figure 3(a) is drawn under the assumption that

$$k_c^1 < k^1 < k^0 < k_c^0. \tag{20a}$$

Figure 3(b), on the other hand, excludes the fixed point of each regime from its domain by assuming that

$$k^1 < k_c^1 < k_c^0 < k^0. \tag{20b}$$

We also draw in Panels (c) and (d) two other cases in which one of the fixed points is within the relevant domain of definition and the other is outside. Intuitively, inequality (20a) holds when the storage technology is much less productive than the neoclassical technology at k^0, while at k^1 banks are able to offer contracts separating high-quality borrowers from low-quality ones. Inequality (20b) on the other hand, means that the storage technology is only a little less productive than the neoclassical technology at k^0, and banks are completely unable to offer contracts permitting them to distinguish low- from high-quality borrowers at k^1.

Dynamic equilibria are solutions to the discontinuous, set-valued difference equation represented by the solid lines in each panel. To ensure that equilibria exist we require that the critical values (k_c^0, k_c^1) satisfy

$$k_c^1 < k_c^0. \tag{21}$$

Otherwise the solid graph in Panel (b) will contain a hole, and a deterministic k_{t+1} will be undefined[12] if k_t were to lie in the interval $[k_c^0, k_c^1]$. Azariadis and Smith (1996b) show that a necessary and sufficient condition for (21) is

$$\gamma\bar{k} < (1 - \gamma)\hat{k}, \tag{22}$$

where ($\hat{k}$, $\bar{k}$) are defined in equations (16a) and (18a).

---

[12] A stochastic equilibrium may exist for k_{t+1} even if the deterministic map is undefined in some region. The solution may require mixed loan strategies from banks, e.g., stochastic credit rationing.

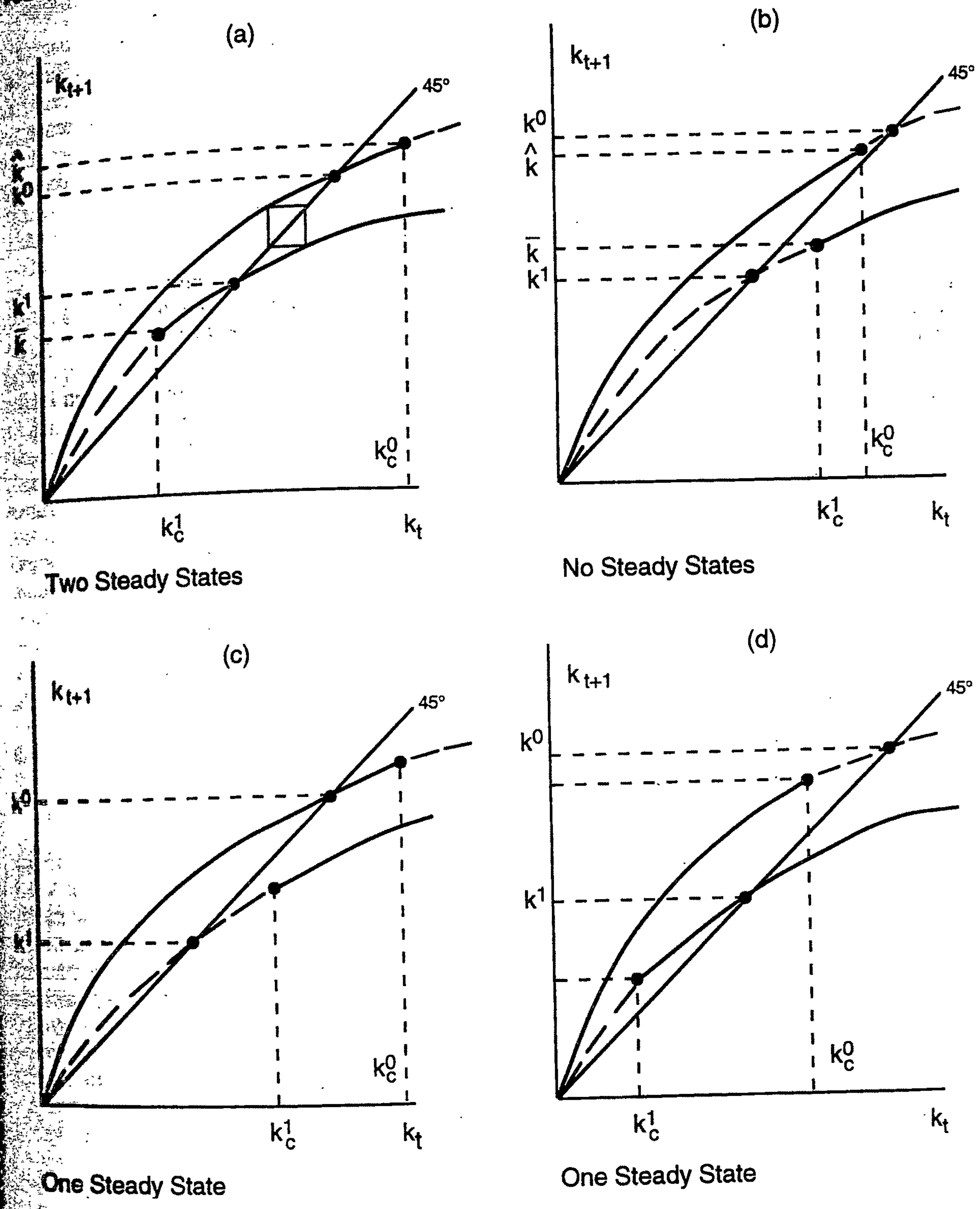


FIGURE 3. REGIME TRANSITIONS

Given the assumption that (22) holds, it is easy to see that each economy in Figure 3 has an *invariant set,* that is, a subset in its state space which traps all solution sequences that start in it. That set is the interval $[k^1, k^0]$ in Figure 3(a), and the interval $(\bar{k}, \hat{k})$ in Figure 3(b). Examples of sequences trapped within the former invariant set are regime 0 equilibria converging to k^0 from below, regime 1 equilibria converging to k^1 from above, as well as the periodic two-cycle depicted in Panel (a).

As we will show, in economies having the configuration of Panel (a), there is in fact a very large set of equilibria displaying deterministic cycles. Hence both indeterminacy of equilibrium and undamped fluctuations are a very real possibility. In economies having the configuration of Panel (b), there are no monotone equilibrium sequences $(\bar{K}_t)$, because all such sequences eventually violate one of the bounds in equation (16d) or (19d). Thus all equilibrium sequences must display transitions between regimes 0 and 1, and therefore all economies having the configuration in Figure 3(b) *necessarily* exhibit fluctuations *despite* the absence of any variations in economic fundamentals.

How do these regime transitions occur? Suppose, for example, that at $t-1$ the economy is in regime 0. Suppose further that savers at t pessimistically believe that the rate of return to savings will be low, or—in other words—that $R_{t+1} = a$ will hold. These low interest rate expectations imply that it is not feasible to have a Walrasian allocation at t, so that banks must ration credit. At the same time, the low rate of return on deposits persuades savers to transfer some savings out of the banking system, and into unintermediated storage. This "disintermediation" validates the necessity of credit rationing. Moreover, the existence of credit rationing breaks the link between the marginal product of capital and the equilibrium rate of interest: this makes it possible for the low returns that depositors expect to actually be observed in equilibrium. Thus a transition between regimes 0 and 1 is associated with a self-fulfilling prophecy of low interest rates, disintermediation, and credit restrictions. Transitions between regimes 1 and 0 occur similarly; if depositors expect that $R_{t+1} > a$ will hold, all savings will be channelled through the banking system, ruling out credit rationing. The equilibrium allocation must now be Walrasian. In short, then, switches between regimes 1 and 0 are accompanied by a self-fulfilling prophecy of rising interest rates and an availability of funds that allows the credit market to clear.

Figure 3(a) depicts a two-period cycle in which regime transitions occur despite the fact that economic fundamentals are consistent with a continuation of the present regime at each date. In Figure 3(b), however, regime transitions *must* occur periodically, for reasons that we now describe.

As we have already argued, a Walrasian allocation is consistent with self-selection in the credit market only if the real rate of interest is sufficiently high. (Excessively low real interest rates provide too little incentive to type 1 agents to work and save when young.) In Figure 3(b), an indefinite continuation in regime 0 leads ultimately to a capital stock which is too high—and a real interest rate which is too low—to be consistent with a Walrasian equilibrium allocation of resources. Thus an economy that is currently in regime 0 must eventually transit into regime 1. By the same token, an economy that remains in regime 1 long enough will ultimately have a capital stock that is so low—and a marginal product of capital that is so high—that lenders have an incentive to pool all borrowing agents, so that self-selection can no longer be sustained. At this point the economy cannot continue in regime 1, and must switch to regime 0. In Figure 3(b), these transitions must clearly repeat themselves. Economic fluctuations associated with these regime transitions are not only possible, but actually inevitable.[13]

In Figure 3(b), then, the operation of the credit market results in the existence of what we will term *reflective barriers* for the economy. An upper bound, or *ceiling*, on real activity exists because excessively low interest rates are inconsistent with a continuation of a Walrasian equilibrium. And a lower bound, or *floor*, on real activity arises from the fact that excessively high marginal products of capital do not present intermediaries with any incentive to ration credit. When these reflective barriers satisfy (20b), real activity will necessarily fluctuate even if economic fundamentals do not vary over time. In these fluctuations, downturns (upturns) are associated with falling (rising) real interest rates, with increased (reduced) rationing of credit, and with resources leaving (reentering) the banking system.

---

[13] A more realistic interpretation of this process, and one that would require a richer menu of borrowers than the one we are using, is to focus on the quality dispersion of loan applicants through the business cycle. In a business expansion, poorly managed firms expand along with better firms, and poor investment projects become marginally profitable. The resulting high dispersion in the quality of loan applications raises the riskiness of intermediary loan portfolios. Banks eventually respond by rationing credit because they lack accurate information about the risk characteristics of individual projects.

As this discussion suggests, it is important whether or not the positive steady states (k^1, k^0) of the two regimes fall outside the interval (k_c^1, k_c^0) defined by the reflective barriers. Accordingly, we arrange our investigation around reflective barriers: we study economies with no binding barriers in Section III, and economies with two binding barriers in Section IV. These discussions will also illustrate what would happen in economies with the configurations depicted in Figures 3(c) and 3(d).

III. Economies with Steady-State Equilibria

When equation (20a) is satisfied, equilibrium sequences (k_t) evolve according to the set-valued difference equation

$$(23)\quad k_{t+1} = \gamma w(k_t); \qquad k_t < k_c^1$$

$$k_{t+1} \in \{\gamma w(k_t), (1-\gamma)w(k_t)\}; \qquad k_c^1 \leq k_t \leq k_c^0;$$

$$(24)\quad k_{t+1} = (1-\gamma)w(k_t); \qquad k_t > k_c^0.$$

We are particularly interested in periodic solutions, that is, in fixed points of iterated maps derived by repeated application of (23) and (24) to describe how today's state variable is related to its value $n = 1, 2, 3, \ldots$, periods hence. Because the map is set valued, iterates of a given order n depend very much on which branch is chosen at each iteration. For instance, always choosing the lower branch $(1-\gamma)w$ of the map will produce a smaller iterate than if we always choose the upper branch γw.

To simplify the mathematical structure, we endow the economy with a logarithmic production function which reduces the maps in (23) and (24) to a piecewise linear difference equation in the logarithm of the capital stock. We assume, in particular, that

$$(25)\qquad f(k) = (A/\theta)k^\theta;$$

$$A > 0, \qquad \theta \in (0, 1).$$

Defining $x = \log k$, we may rewrite equations (23) and (24) in their linear form:

$$(26)\quad x_{t+1} = f_0(x_t) \equiv \theta(x_t - x^0) + x^0;$$

$$x_t < x_c^1$$

$$x_{t+1} \in \{f_0(x_t), f_1(x_t)\}; \qquad x_c^1 \leq x_t \leq x_c^0$$

$$x_{t+1} = f_1(x_t) \equiv \theta(x_t - x^1) + x^1; \qquad x_t > x_c^0,$$

where

$$(27)\quad x^0 = [1/(1-\theta)]\log[\gamma(1-\theta)A/\theta]$$

$$\equiv \log k^0;$$

$$(28)\quad x^1 = [1/(1-\theta)]$$

$$\times \log[(1-\gamma)(1-\theta)A/\theta]$$

$$\equiv \log k^1$$

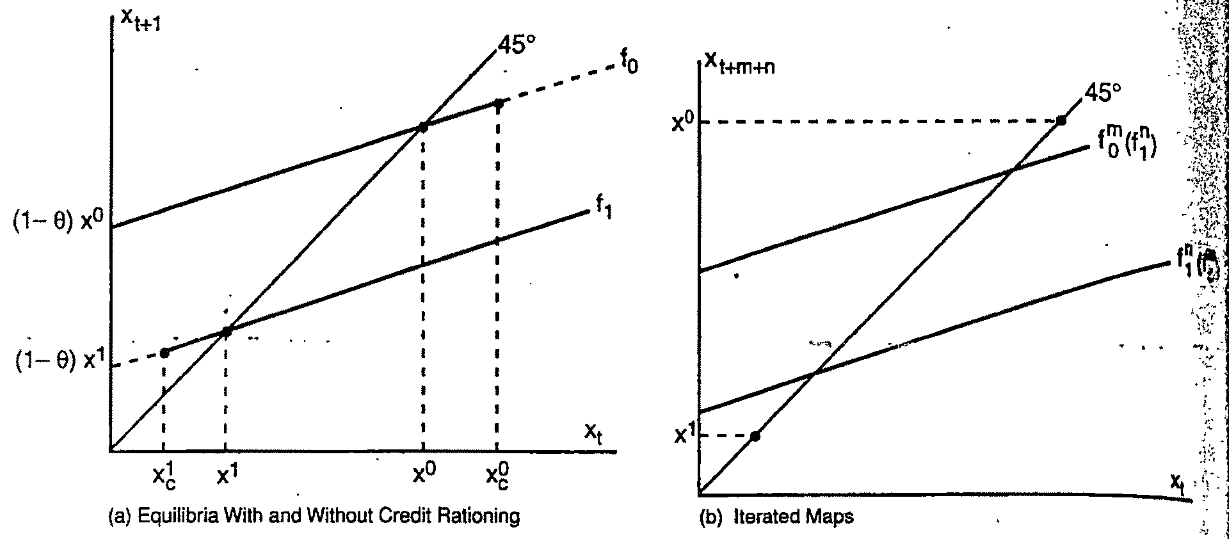
are the fixed points of regimes 0 and 1, respectively. It is also useful to compute the critical points for the two regimes. Under the technological assumption in equation (25), these are given by:

$$(29\text{a})\quad x_c^0 = (1/\theta)\log\left\{\left[\frac{\theta}{\gamma(1-\theta)A}\right] \times \left[\frac{A(1-\gamma)}{a\gamma}\right]^{1/(1-\theta)}\right\}$$

$$\equiv \log k_c^0;$$

$$(29\text{b})\quad x_c^1 = (1/\theta)\log\left\{\left[\frac{\theta}{(1-\gamma)(1-\theta)A}\right] \times \left[\frac{A(1-\gamma)}{a}\right]^{1/(1-\theta)}\right\}$$

$$\equiv \log k_c^1.$$



(a) Equilibria With and Without Credit Rationing

(b) Iterated Maps

FIGURE 4. EQUILIBRIUM LAW OF MOTION
(COBB-DOUGLAS PRODUCTION)

The map defined by (26) is depicted in Figure 4(a). In order to ensure that this map contains no holes and that deterministic equilibria exist, we assume $x_c^1 < x_c^0$ or, equivalently,

$$(30) \qquad 1 > \gamma\left(\frac{\gamma}{1-\gamma}\right)^{1-\theta}.$$

In addition, capital dominates storage in rate of return for all $k \leq k^0$ iff $f'(k^0) > a$ holds. This condition proves to be equivalent to the restriction

$$(31) \qquad \gamma < \frac{\theta/(1-\theta)}{\max\{0, a\}}.$$

Finally, an economy has no effective reflective barriers if the map (26) contains in its domain of definition the fixed points (x^0, x^1) of each regime. This inclusion means that

$$(32) \qquad x_c^1 < x^1 < x^0 < x_c^0$$

or, equivalently, that

$$(33) \quad a\gamma^2/(1-\gamma) < \theta/(1-\theta) < a.$$

In the remainder of this section we study solutions to equation (26), maintaining inequalities (30)–(33) as restrictions on the economy's parameter space (a, γ, θ, A). Inequality (32), in particular, implies that the steady state of each regime is an equilibrium and, hence, regime transitions *need not* occur in equilibrium. If there are switches in regimes, this will be due to self-confirming expectations by depositors about the behavior of intermediaries and the yields on bank liabilities.

We are now prepared to establish the main result of this section, which indicates that this economy can display very strong indeterminacies and excessive fluctuations in perfect-foresight equilibria. It is an almost immediate implication that there is also a wide variety of equilibria where the economy experiences stochastic shifts between regimes 0 and 1 in a Markovian manner, with the probability of regime transitions depending potentially on time, history, or the state of the system.

To begin, it will be useful to orient our formal discussion around the notion of "an (m, n) cycle." Loosely speaking, an (m, n) cycle is defined to be a deterministic cycle of m periods spent in regime 0, followed by n periods spent in regime 1 (and so on). In order to define an (m, n) cycle more formally, we begin by considering compound iterates of the maps f_0 and f_1 introduced in equation (26). In particular, let $f_0^m(f_1^n)$ denote the map $f_0(f_1)$ iterated with itself $m(n)$ times. Then:

$$(34a) \quad f_0^m(x) = \theta^m x + (1-\theta^m)x^0;$$

$$(34b) \quad f_1^n(x) = \theta^n x + (1-\theta^n)x^1.$$

Compound iterates of the form $f_0^m[f_1^n(x)]$ and $(f_1^n[f_0^m(x)])$ describe orbits that spend n periods in regime 1 (m periods in regime 0) followed by m periods in regime 0 (n periods in regime 1). For given (m, n), these iterated maps satisfy

$$f_0^m[f_1^n(x)] = \theta^{m+n}x + (1-\theta^m)x^0 + \theta^m(1-\theta^n)x^1; \tag{34c}$$

$$\begin{aligned} f_1^n[f_0^m(x)] &= \theta^{m+n}x + \theta^n(1-\theta^m)x^0 \\ &\quad + (1-\theta^n)x^1. \end{aligned} \tag{34d}$$

Figure 4(b) graphs these maps. Since $x^0 > x^1$, we have

$$f_0^m(f_1^n) > f_1^n(f_0^m) \qquad \forall x. \tag{35}$$

Finally, we may compute two types of double compound iterates that change regime twice. One starts with $m - q$ periods in regime 0, continues with n periods in regime 1, and follows with q periods in regime 0 (and so on). The second starts with $n - p$ periods in regime 1, followed by m periods in regime 0, and ending with p periods in regime 1 again (and so on). Maps that describe such orbits are given by

$$\begin{aligned} f_0^q[f_1^n(f_0^{m-q}(x))] &= \theta^{m+n}x + [\theta^{n+q}(1-\theta^{m-q}) \\ &\quad + (1-\theta^q)]x^0 \\ &\quad + \theta^q(1-\theta^n)x^1 \end{aligned} \tag{36a}$$

and

$$\begin{aligned} f_1^p[f_0^m(f_1^{n-p}(x))] &= \theta^{m+n}x + [\theta^{m+p}(1-\theta^{n-p}) \\ &\quad + (1-\theta^p)]x^1 \\ &\quad + \theta^p(1-\theta^m)x^0, \end{aligned} \tag{36b}$$

respectively. Then we have the following proposition.

Definition: For given $m \geq 1$ and $n \geq 1$, the sequence $(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m, \tilde{x}_{m+1}, \dots, \tilde{x}_{m+n})$ constitutes an (m, n) cycle if

(i) $\tilde{x}_i \in [x^1, x^0], \quad \forall i = 1, \dots, m+n,$ and either
(ii) $\tilde{x}_i = f_0^{i-1} \circ f_1^n \circ f_0^{m+1-i}(\tilde{x}_i), \quad i = 1, \dots, m,$[14] or
(iii) $\tilde{x}_i = f_1^{i-m-1} \circ f_0^m \circ f_1^{n+m+1-i}(\tilde{x}_i)$ for $i = m+1, \dots, m+n$.

Thus the values $\tilde{x}_i$ are each fixed points of the appropriate compound maps.

We are now prepared to state the following proposition. Its proof appears in Azariadis and Smith (1996b).

PROPOSITION 3: *(a) For all positive integers m and n, there exists an asymptotically stable (m, n) cycle; (b) Any (m, n) cycle has the property that* $\tilde{x}_1 \leq \tilde{x}_i \leq \tilde{x}_{m+1}, \quad \forall i = 1, \dots, m+1$. *Moreover,*

$$\tilde{x}_1 = [1 - \Gamma(m,n)]x^0 + \Gamma(m,n)x^1; \tag{37}$$

$$\tilde{x}_{m+1} = [1 - \theta^m\Gamma(m,n)]x^0 + \theta^m\Gamma(m,n)x^1, \tag{38}$$

where

$$\Gamma(m,n) \equiv (1-\theta^n)/(1-\theta^{m+n}). \tag{39}$$

These equations imply that, for any (m, n), both of the extreme periodic points $\tilde{x}_1$ and $\tilde{x}_{m+1}$ lie within the invariant interval $[x^0, x^1]$; hence the remaining fixed points $(\tilde{x}_2, \dots, \tilde{x}_{m-1}, \tilde{x}_m, \dots, \tilde{x}_{m+n})$ will also lie in that interval. Proposition 3 specifies the sense in which this economy displays a high-dimensional indeterminacy, and in which there is a wide variety of perfect-foresight equilibria that display undamped oscillations. In addition, these equilibria can be attained starting from a variety of initial capital stocks. The proposition also demonstrates how private information in the credit market spawns indeterminate equilibria in an overlapping generations economy which otherwise lacks all the features we normally hold responsible for indeterminacy:[15] there is one sector, there are no paper assets, no large income effects, no nonconvexities in production or preferences, no monopolistic competition, nor are there any other significant departures from the key assumptions of Arrow and Debreu—except private information.

---

[14] Here f^0 denotes the identity map.

It is also straightforward to construct equilibria in which transitions between the two regimes occur stochastically. In particular, let $\Phi_t \in \{0, 1\}$ index the regime, and let

$$(40a)\qquad \pi^t_{00} = \text{prob}(\Phi_{t+1} = 0 \mid \Phi_t = 0);$$

$$(40b)\qquad \pi^t_{11} = \text{prob}(\Phi_{t+1} = 1 \mid \Phi_t = 1).$$

Then the sequence of matrices

$$(41)\qquad \Pi_t = \begin{pmatrix} \pi^t_{00} & 1 - \pi^t_{00} \\ 1 - \pi^t_{11} & \pi^t_{11} \end{pmatrix}$$

induces a (possibly nonstationary) Markov process on the regime index Φ_t. Moreover, competitive equilibria exist for arbitrary sequences (π^t_{00}, π^t_{11}), and the elements of these sequences can depend in essentially any fashion on the history of the economy.

To summarize, in an economy where the steady-state equilibria of each regime lie within the economy's reflective barriers, the expectations of depositors become paramount in determining the type of equilibrium that will be observed. In particular, depositors must form expectations about yields on bank liabilities. Optimistic expectations of high yields lead to Walrasian regimes without credit rationing, and the absence of rationing validates depositors' expectations. Pessimistic expectations of low yields lead to regimes of disintermediation, which forces banks to ration credit. This credit rationing again validates depositors' beliefs. Periodic cycles correspond to alternating, self-confirming waves of optimism and pessimism.

In Section IV, we consider economies where the steady-state equilibria of the two regimes lie outside the economy's reflective barriers. When this occurs, the existence of these barriers limits the scope of depositors' expectations by ruling out persistently optimistic or persistently pessimistic rational expectations equilibria. The result is that *all competitive equilibria must display endogenous volatility,* and that this volatility *cannot* vanish asymptotically.

IV. Binding Reflective Barriers

We next describe equilibria when neither steady state lies within the economy's reflective barriers. Therefore, in this section we assume that

$$(42)\qquad x^1 < x^1_c < x^0_c < x^0;$$

Evidently, then, neither steady state constitutes a legitimate competitive equilibrium, nor do any monotonic sequences (x_t) that converge to one of the steady states. We now describe what kinds of equilibria can be observed here.

In order to do so, we define two values, $\bar{x}_F$ and $\bar{x}_c$, by the relations

$$\bar{x}_F \equiv \theta x^1_c + (1 - \theta)x^1 \equiv f_1(x^1_c);$$

$$\bar{x}_c \equiv \theta x^0_c + (1 - \theta)x^0 \equiv f_0(x^0_c).$$

It follows from these definitions that $\bar{x}_F$ is simply the point that succeeds x^1_c in regime 1, while $\bar{x}_c$ is the point that succeeds x^0_c in regime 0. These two points represent a floor or ceiling, respectively, on the value of x_t that can be attained in each regime. From equations (28) and (29) it is possible to obtain the following closed-form expressions for $\bar{x}_F$ and $\bar{x}_c$:

$$(43a)\qquad \bar{x}_F = \frac{1}{1-\theta}\log\left[\frac{A(1-\gamma)}{a}\right];$$

$$(43b)\qquad \bar{x}_C = \frac{1}{1-\theta}\log\left[\frac{A(1-\gamma)}{a\gamma}\right].$$

---

[15] See Azariadis (1993 Ch. 26), Roger E. A. Farmer (1993), and Roger Guesnerie and Michael Woodford (1993) for recent surveys of indeterminate equilibria.

An examination of Figure 3(b) should now convince the reader of the following proposition.

PROPOSITION 4: *(i) Every equilibrium sequence of this economy has an upper and a lower turning point, and (ii) both turning points lie inside the invariant set $[\bar{x}_F, \bar{x}_c]$, which attracts all equilibrium sequences in finite time.*

In view of the proposition, we now focus our attention on the potential existence of equilibria displaying (m, n) cycles. Such equilibria are defined exactly as in Section III, except that the invariant interval is now $[\bar{x}_F, \bar{x}_c]$ rather than $[x^1, x^0]$.

As in Section III, let $\tilde{x}_1$ and $\tilde{x}_{m+1}$ denote the (unique) fixed points of the compound maps $f_0^m(f_1^n)$ and $f_1^n(f_0^m)$, respectively. Then it is immediately apparent that an (m, n) cycle exists, for fixed $m \geq 1$ and $n \geq 1$, if and only if

$$\bar{x}_F \leq \tilde{x}_1 < \tilde{x}_{m+1} \leq \bar{x}_c. \tag{44}$$

The next proposition states a necessary condition for (44) to hold for some combination (m, n).

PROPOSITION 5: *Define the parameters A_0, A_1, and z by the following relations:*[16]

$$A_0 \equiv (\bar{x}_c - x^1) / (x^0 - x^1); \tag{45a}$$

$$A_1 \equiv (\bar{x}_F - x^1) / (x^0 - x^1); \tag{45b}$$

$$z \equiv \max\{A_1 / (1 - A_1), (1 - A_0) / A_0\}. \tag{45c}$$

Then a deterministic (m, n) cycle exists, for some integers $m \geq 1$ and $n \geq 1$, if $\theta \geq z$.

The proof of Proposition 5 appears in Azariadis and Smith (1996b). The proposition establishes a sufficient condition for the existence of an (m, n) cycle.

When $m > (<) n$ holds, an (m, n) cycle is asymmetric in the sense that the number of periods of expansion, m, exceeds (is less than) the number of periods of contraction, n. When $m = n$, on the other hand, we have a symmetric cycle. We now show that the existence of certain asymmetric (m, n) cycles always implies the existence of at least one—and typically many—symmetric cycles.

PROPOSITION 6: *(a) Suppose that $A_0 + A_1 \leq 1$ holds. Then the existence of an (m, n) cycle with $m > n$ implies the existence of an (n, n) cycle. Moreover, if k is any integer with $k \leq n$, there exists a (k, k) cycle. (b) Suppose that $A_0 + A_1 \geq 1$ holds. Then the existence of an (m, n) cycle with $n > m$ implies the existence of an (m, m) cycle. Moreover, if k is any integer with $k \leq m$, there exists a (k, k) cycle.*

Proposition 6 is proved in Azariadis and Smith (1996b). It states conditions under which the existence of certain (m, n) cycles implies the existence of other (symmetric) cycles. Thus a variety of indeterminacies continues to be observed when these conditions are satisfied.

A. *Maximally Persistent Cycles*

The cycles discussed in Section III have the property that there are periodic changes in the beliefs of depositors about regimes, and that these belief changes trigger self-fulfilling regime transitions. However, nothing in Section III necessitates these changes in beliefs; in some sense, then, their occurrence may depend on arbitrary historical events. In economies satisfying (42), on the other hand, objective conditions do not permit the beliefs of depositors to remain constant indefinitely. These beliefs must undergo repeated transitions, as must the regime that governs equilibrium behavior. We might ask, then, about the properties of periodic equilibria where depositor beliefs are, loosely speaking, revised as infrequently as possible, and hence where regime switching occurs as infrequently as possible. Our interest in such equilibria leads us to now turn our attention to the existence of what we term *maximally persistent* (m, n) cycles. In particular, a maximally persistent (m, n) cycle has the property that, if there is also a (p, q) cycle, $p + q \leq m + n$ holds. Obviously, a maximally persistent (m, n) cycle exists if (42) holds, and $\theta \geq z$.

---

[16] Equation (42) implies that $0 < A_1 < A_0 < 1$ holds. The parameters A_0 and A_1 measure, respectively, the relative distance of the ceiling and floor from the nearest steady state and, hence, reveal how tight reflective barriers [illegible] As $A_0 \to 1$ and $A_1 \to 0$, floors and ceilings become less [illegible] permitting equilibria to approach the steady states [illegible] more closely before reversing direction.

Maximally persistent cycles are of interest for two reasons. First, as we have said, these cycles correspond to equilibria with the least volatile depositor expectations that can be observed. Second, if (m, n) is a cycle of maximal persistence, and if $m > (<) n$, all cycles $(m - k, n)$ $[(m, n - k)]$ exist for $k = 1, \ldots, m - n$ $(k = 1, \ldots, n - m)$. Such cycles, of course, exhibit depositor expectations of greater volatility than the (m, n) cycle.

We now state a formal proposition describing maximally persistent cycles. Its proof appears in Azariadis and Smith (1996b).

PROPOSITION 7: *(a) Suppose that inequality (42), $\theta \geq z$, and $A_0 + A_1 \geq 1$ are all satisfied. Let $\hat{s} \geq 1$ denote the largest integer solution to $\theta^s \geq A_1/(1 - A_1)$, and let $\hat{m} \geq 1$ denote the largest integer solution to $\theta^m \geq (1 - A_0)/(1 - A_0\theta^{\hat{s}})$. Then $\hat{m} \geq \hat{s}$ holds, and $(\hat{m}, \hat{s})$ is the maximally persistent (m, n) cycle. (b) Suppose now that (42), $\theta \geq z$, and $A_0 + A_1 \leq 1$ hold. Let $\hat{s} \geq 1$ denote the largest integer solution to $\theta^s \geq (1 - A_0)/A_0$, and let $\hat{n} \geq 1$ denote the largest integer solution to $\theta^n \geq A_1/[1 - (1 - A_1)\theta^{\hat{s}}]$. Then $\hat{n} \geq \hat{s}$ holds, and $(\hat{s}, \hat{n})$ is the maximally persistent (m, n) cycle.*

Note that the proposition asserts that maximally persistent cycles have longer expansions than contractions when $A_0 + A_1 > 1$, because x^0 is closer to the ceiling on real activity than x^1 is to the floor. Maximally persistent cycles have longer contractions than expansions in the opposite case, and for precisely the opposite reason.

Prior to concluding this section, it is appropriate to mention that (m, n) cycles by no means exhaust the set of possible deterministic periodic equilibria. For example, there may be equilibria with m periods of expansion followed by n periods of contraction, p periods of expansion again, and then q periods of contraction (and so on), with $m \neq p$ and $n \neq q$. Indeed, such equilibria may exist even if $\theta < z$ holds. In addition, stochastic equilibria of the type described in Section III will typically exist. What does *not* exist here are equilibrium sequences that converge to steady states.

V. CONCLUSIONS

A long tradition in monetary economics holds that the unrestricted operation of financial markets can lead to indeterminacies and excessive fluctuations relative to fundamentals. Previous models that deliver this result, however, have typically depended on the existence of nominal assets [models with nominal assets and production include Ben S. Bernanke and Gertler (1989), Azariadis and Smith (1996a), John H. Boyd and Smith (1997), or Schreft and Smith (1997)], large income effects [Azariadis (1981); Azariadis and Guesnerie (1986)], nonconvexities [Jess Benhabib and Farmer (1994)], monopolistic competition [Jordi Galí (1994); Russell Cooper and Jaoa Ejarque (1994)], or multiple sectors [Nobuhiro Kiyotaki and John Moore (1993); Bencivenga and Smith (1997)]. We have examined a one-sector neoclassical growth model with none of these features, but in which capital investment must be credit financed, and in which credit markets are characterized by the presence of an adverse selection problem. In this model we have shown that the existence of two equilibrium regimes is possible: a Walrasian regime in which incentive constraints are nonbinding, and a regime of credit rationing in which the rationing is necessary to induce self-selection in loan markets. For some range of current capital stocks either regime is consistent with the existence of an equilibrium; therefore, it is possible to observe equilibria in which deterministic or stochastic transitions occur between regimes. Many of these equilibria will display oscillations that do not die out. Moreover, for some configurations of parameters (those considered in Section IV), the *only* equilibria that can be observed display fluctuations that do not vanish asymptotically. In such economies excessive fluctuations are not only possible, but indeed are a necessary feature of any equilibrium.

Relative to much of the existing literature on indeterminancies and endogenous volatility, our analysis contains several new features

One is the potential for either deterministic or stochastic regime switching. This potential derives from the fact that dynamic equilibria in our model satisfy an unusual, set-valued, discontinuous difference equation which contains two well-defined, partially overlapping domains with distinct structural regimes. At intermediate levels of economic activity transitions between these regimes are possible. Indeed, the resulting equilibria may be indexed by a stochastic process that governs an unobserved regime-switching variable in the manner proposed by James D. Hamilton (1989, 1990). The economic interpretation of this state variable is to regard it as an index of savers' expectations about credit market conditions. An expanding body of evidence suggests that such nonlinear mechanisms accord well with the behavior of many aggregate time series.[17] Yet, while regime-switching models have played an important role in an array of empirical investigations, there has so far been little theoretical work providing any underpinning for them. In addition, much of the theoretical literature studying endogenous volatility has the feature that there are a large set of equilibria, only some of which exhibit fluctuations. We have demonstrated the possibility that such fluctuations not only *can* be observed, but *must* be observed under certain conditions. When these conditions obtain, economic volatility that stems from changing credit market conditions cannot be avoided.

REFERENCES

Azariadis, Costas. "Self-Fulfilling Prophecies." *Journal of Economic Theory*, December 1981, *25*(2), pp. 380–96.

——. *Intertemporal macroeconomics*. Oxford: Blackwell, 1993.

Azariadis, Costas and Guesnerie, Roger. "Sunspots and Cycles." *Review of Economic Studies*, October 1986, *53*(4), pp. 725–37.

Azariadis, Costas and Smith, Bruce. "Private Information, Money and Growth: Indeterminacy, Fluctuations, and the Mundell-Tobin Effect." *Journal of Economic Growth*, September 1996a, *1*(3), pp. 309–32.

——. "Appendix to Financial Intermediation and Regime Switching in Business Cycles." Mimeo, University of California, Los Angeles, 1996b.

Beaudry, Paul and Koop, Gary. "Do Recessions Permanently Change Output?" *Journal of Monetary Economics*, April 1993, *31*(2), pp. 149–63.

Bencivenga, Valerie R. and Smith, Bruce. "Some Consequences of Credit Rationing in an Endogenous Growth Model." *Journal of Economic Dynamics and Control*, January 1993, *17*(1), pp. 97–112.

——. "Unemployment, Migration, and Growth." *Journal of Political Economy*, June 1997, *105*(3), pp. 582–608.

Benhabib, Jess and Farmer, Roger E. A. "Indeterminacy and Increasing Returns." *Journal of Economic Theory*, June 1994, *63*(1), pp. 19–41.

Bernanke, Ben S. and Gertler, Mark. "Agency Costs, Net Worth and Business Fluctuations." *American Economic Review*, March 1989, *79*(1), pp. 14–31.

Boldin, M. "Characterizing Business Cycles with a Markov Switching Model." Working paper, Federal Reserve Bank of New York, 1990.

Boyd, John H. and Smith, Bruce D. "Credit Market Imperfections in a Monetary Growth Model." *Journal of Economic Theory*, 1998, *11*(2), pp. 241–73.

Cooper, Russell and Ejarque, Jaoa. "Financial Intermediation and Aggregate Fluctuations: A Quantitative Analysis." Mimeo, Boston University, 1994.

Diamond, Peter A. "National Debt in a Neoclassical Growth Model." *American Economic Review*, December 1965, *55*(5), pp. 1026–50.

Diebold, Francis and Rudebusch, Glenn. "Measuring Business Cycles: A Modern Perspective." Mimeo, University of Pennsylvania, 1994.

Farmer, Roger E. A. *The economics of self-fulfilling prophesies*. Cambridge, MA: MIT Press, 1993.

Friedman, Milton. *A program for monetary stability*. New York: Fordham University Press, 1960.

---

[17] See, for instance, Salih Neftçi (1984), M. Boldin (1990), Paul Beaudry and Gary Koop (1993), Francis Diebold and Glenn Rudebusch (1994), and Simon Potter

Friedman, Milton and Schwartz, Anna J. *A monetary history of the United States, 1867–1960*. Princeton, NJ: Princeton University Press, 1963.

Galí, Jordi. "Multiple Equilibria in a Growth Model with Monopolistic Competition." *Journal of Economic Theory*, June 1994, *63*(1), pp. 73–96.

Gertler, Mark and Hubbard, R. Glenn. "Financial Factors in Business Fluctuations." Financial Market Volatility: Federal Reserve Bank of Kansas City Symposium Series, August 17–19, 1988.

Gertler, Mark and Rogoff, Kenneth. "North-South Lending and Endogenous Domestic Capital Market Inefficiencies." *Journal of Monetary Economics*, October 1990, *26*(2), pp. 245–66.

Guesnerie, Roger and Woodford, Michael. "Endogenous Fluctuations." Mimeo, DELTA, 1993.

Hamilton, James D. "A New Approach to the Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica*, March 1989, *57*(2), pp. 357–84.

———. "Analysis of Time Series Subject to Changes in Regime." *Journal of Econometrics*, July 1990, *45*(1), pp. 39–70.

Keynes, John Maynard. *The general theory of employment, interest and prices*. London: Macmillan, 1936.

Kiyotaki, Nobuhiro and Moore, John. "Credit Cycles." Mimeo, University of Minnesota, 1993.

Kreps, David. *A course in microeconomic ory*. Princeton, NJ: Princeton Univer Press, 1990.

Neftçi, Salih. "Are Economic Time Se Asymmetric over the Business Cycle *Journal of Political Economy*, April 19 *92*(2), pp. 307–28.

Potter, Simon. "A Nonlinear Approach to US GNP." Mimeo, University of Califor Los Angeles, 1994.

Rothschild, Michael and Stiglitz, Joseph. "Eq librium in Competitive Insurance Market *Quarterly Journal of Economics*, Nove ber 1976, *90*(4), pp. 629–49.

Schreft, Stacey L. "Credit Controls: 1980 *Federal Reserve Bank of Richmond Eco nomic Review*, November–December 199 *76*(6), pp. 25–55.

Schreft, Stacey L. and Owens, Raymond F. "Id tifying Credit Crunches." *Contempora Economic Policy*, April 1995, *13*(2), 63–76.

Schreft, Stacey L. and Smith, Bruce. "Mon Banking and Capital Formation." *Journ of Economic Theory*, March 1997, *73*(1), pp. 157–82.

Simons, Henry. *Monetary policy for a free ciety*. Chicago: University of Chic Press, 1948.

Stiglitz, Joseph and Weiss, Andrew. "Credit Ra tioning in Models with Asymmetric Info mation." *American Economic Review*, Ju 1981, *71*(3), pp. 393–410.

Stock Markets, Banks, and Economic Growth

By Ross Levine and Sara Zervos*

Do well-functioning stock markets and banks promote long-run economic growth? This paper shows that stock market liquidity and banking development both positively predict growth, capital accumulation, and productivity improvements when entered together in regressions, even after controlling for economic and political factors. The results are consistent with the views that financial markets provide important services for growth, and that stock markets provide different services from banks. The paper also finds that stock market size, volatility, and international integration are not robustly linked with growth, and that none of the financial indicators is closely associated with private saving rates. (*JEL* G00, O16, F36)

Considerable debate exists on the relationships between the financial system and economic growth. Historically, economists have focused on banks. Walter Bagehot (1873) and Joseph A. Schumpeter (1912) emphasize the critical importance of the banking system in economic growth and highlight circumstances when banks can actively spur innovation and future growth by identifying and funding productive investments. In contrast, Robert E. Lucas, Jr. (1988) states that economists "badly over-stress" the role of the financial system, and Joan Robinson (1952) argues that banks respond passively to economic growth. Empirically, Robert G. King and Levine (1993a) show that the level of financial intermediation is a good predictor of long-run rates of economic growth, capital accumulation, and productivity improvements.

Besides the historical focus on banking, there is an expanding theoretical literature on the links between stock markets and long-run growth, but very little empirical evidence. Levine (1991) and Valerie R. Bencivenga et al. (1995) derive models where more liquid stock markets—markets where it is less expensive to trade equities—reduce the disincentives to investing in long-duration projects because investors can easily sell their stake in the project if they need their savings before the project matures. Enhanced liquidity, therefore, facilitates investment in longer-run, higher-return projects that boost productivity growth. Similarly, Michael B. Devereux and Gregor W. Smith (1994) and Maurice Obstfeld (1994) show that greater international risk sharing through internationally integrated stock markets induces a portfolio shift from safe, low-return investments to high-return investments, thereby accelerating productivity growth. These liquidity and risk models, however, also imply that greater liquidity and international capital market integration ambiguously affect saving rates. In fact, higher returns and better risk sharing may induce saving rates to fall enough such that overall growth slows with more liquid and internationally integrated financial markets. Moreover, theoretical debate exists about whether greater stock market liquidity actually encourages a shift to higher-return projects that stimulate productivity growth. Since more liquidity makes it easier to sell shares, some

---

* Levine: Department of Economics, University of Virginia, Charlottesville, VA 22903; Zervos: Barclay's Capital, Canary Wharf, London, U.K. We thank Mark Baird, Valerie Bencivenga, John Boyd, Jerry Caprio, Asli Demirgüç-Kunt, Doug Diamond, Bill Easterly, Michael [illegible] Bruce Smith, two anonymous referees, and seminar participants at Arizona State University, Cornell University, Dartmouth College, Harvard Institute for International Development, the University of Virginia, and the University of Washington for helpful comments. We received excellent research assistance from Michelle [illegible] and Ti Caudron. Much of the work on this paper was done while the authors were employed by the World Bank. Opinions expressed are those of the authors and do not necessarily reflect those of the World Bank, its staff, or member countries.

argue that more liquidity reduces the incentives of shareholders to undertake the costly task of monitoring managers (Andrei Shleifer and Robert W. Vishny, 1986; Amar Bhide, 1993). In turn, weaker corporate governance impedes effective resource allocation and slows productivity growth. Thus, theoretical debate persists over the links between economic growth and the functioning of stock markets.[1]

This paper empirically investigates whether measures of stock market liquidity, size, volatility, and integration with world capital markets are robustly correlated with current and future rates of economic growth, capital accumulation, productivity improvements, and saving rates using data on 47 countries from 1976 through 1993. This investigation provides empirical evidence on the major theoretical debates regarding the linkages between stock markets and long-run economic growth. Moreover, we integrate this study into recent cross-country research on financial intermediation and growth by extending the King and Levine (1993a) analysis of banking and growth to include measures of the functioning of stock markets. Specifically, we evaluate whether banking and stock market indicators are *both* robustly correlated with current and future rates of economic growth, capital accumulation, productivity growth, and private saving. If they are, then this suggests that both banks and stock markets have an independent empirical connection with contemporaneous and future long-run growth rates.

We find that stock market liquidity—*as measured both by the value of stock trading relative to the size of the market and by the value of trading relative to the size of the economy*—is positively and significantly correlated with current and future rates of economic growth, capital accumulation, and productivity growth. Stock market liquidity is a robust predictor of real per capita gross domestic product (GDP) growth, physical capital growth, and productivity growth after controlling for initial income, initial investment in education, political stability, fiscal policy, openness to trade, macroeconomic stability, and the forward-looking nature of stock prices. Moreover, the level of banking development—*as measured by bank loans to private enterprises divided by GDP*—also enters these regressions significantly. Banking development and stock market liquidity are both good predictors of economic growth, capital accumulation, and productivity growth. The other stock market indicators do not have a robust link with long-run growth. Volatility is insignificantly correlated with growth in most specifications. Similarly, market size and international integration are not robustly linked with growth, capital accumulation, and productivity improvements. Finally, none of the financial indicators is robustly related to private saving rates.

The results have implications for a variety of theoretical models. The strong, positive connections between stock market liquidity and faster rates of growth, productivity improvements, and capital accumulation confirm Levine's (1991) and Bencivenga et al. (1995) theoretical predictions. We do not find any support, however, for theories that more liquid or more internationally integrated capital markets negatively affect saving and growth rates or that greater liquidity retards productivity growth.[2] Further, the evidence does not support the belief that stock return volatility hinders investment and resource lation (J. Bradford DeLong et al., 1989). Finally, the data also suggest that banks provide different services from those of stock markets. Measures of both banking development and stock market liquidity enter the growth regression significantly. Thus, to understand the relationship between financial systems and economic growth, we need theories in which stock markets and banks arise simultaneously to provide different bundles of financial services.

---

[1] In terms of banks, Douglas W. Diamond (1984), John H. Boyd and Edward C. Prescott (1986), and Stephen D. Williamson (1986) develop models where financial intermediaries—coalitions of agents—lower the costs of obtaining information about firms from what those costs would be in atomistic capital markets where each investor must acquire information individually. Based on these core models, King and Levine (1993b) show that, by lowering information costs, financial intermediaries foster more efficient resource allocation and thereby accelerate technological innovation and long-run growth. Jeremy Greenwood and Boyan Jovanovic (1990) develop a model in which financial intermediaries affect, and are affected by, economic growth. See the review by Levine (1997).

[2] See Bencivenga and Smith (1991) and Obstfeld (1994) for parameter values that lead to lower saving and growth rates with greater liquidity or risk sharing, respectively. The data are inconsistent with these parameter values. Note, however, that these models have parameter values that are consistent with our empirical findings: (a) liquidity is positively associated with economic growth; and (b) neither liquidity nor international capital market integration is associated with private saving

A few points are worth emphasizing in interpreting the results. First, since Levine and David Renelt (1992) show that past researchers have been unable to identify empirical links between growth and macroeconomic indicators that are robust to small changes in the conditioning information set, we check the sensitivity of the results to changes in a large conditioning information set. Stock market liquidity and banking development are positively and robustly correlated with current and future rates of economic growth even after controlling for many other factors associated with economic growth. Second, almost all previous cross-country studies of growth focus on data where both the dependent and explanatory variables are averaged over the entire sample period. Besides examining this contemporaneous relationship, we study whether stock market and banking development measured at the beginning of the period robustly predict future rates of economic growth, capital accumulation, productivity growth, and private saving rates. We find that stock market liquidity and banking development both predict long-run growth, capital accumulation, and productivity improvements. Although this investigation does not establish the direction of causality between financial-sector development and growth, the results show that the strong link between financial development and growth does not merely reflect contemporaneous shocks to both, that stock market and banking development do not simply follow economic growth, and that the predictive content of the financial development indicators does not just represent the forward-looking nature of stock prices. This paper's results are certainly consistent with the view that the services provided by financial institutions and markets are important for long-run growth. Finally, this paper's aggregate cross-country analyses complement recent microeconomic evidence. Asli Demirgüç-Kunt and Vojislav Maksimovic (1996) show that firms in countries with better-functioning banks and equity markets grow faster than predicted by individual firm characteristics, and Raghuram G. Rajan and Luigi Zingales (1998) show that industries that rely more on external finance prosper more in countries with better-developed financial markets.

Raymond Atje and Jovanovic (1993) present a cross-country study of stock markets and economic growth. They find a significant correlation between growth over the period 1980–1988 and the value of stock market trading divided by GDP for 40 countries. We make several contributions. Besides increasing the number of countries by almost 20 percent and almost doubling the number of years in the sample, we construct additional measures of stock market liquidity, a measure of stock return volatility, and two measures of stock market integration in world capital markets and incorporate these measures into our study of stock markets, banks, and economic growth. Furthermore, we control for economic and political factors that may influence growth to gauge the sensitivity of the results to changes in the conditioning information set. Moreover, we control for the potential forward-looking nature of financial prices since we want to gauge whether the functioning of stock markets and banks is tied to economic performance, not whether agents anticipate faster growth. Also, we use the standard cross-country growth regression framework of Robert J. Barro (1991) to make comparisons with other work easier, systematically test for the importance of influential observations, and correct for heteroskedasticity. Finally, besides the direct link with growth, we also study the empirical connections between stock market development and physical capital accumulation, productivity improvements, and private saving rates.

The next section presents measures of stock market and banking development, as well as four *growth indicators*—measures of the rate of economic growth, capital accumulation, productivity growth, and private saving. Section II examines the relationship between the four growth indicators and stock market liquidity, size, volatility, international capital market integration, as well as the level of banking development. Section III concludes.

I. Measuring Stock Market and Banking Development and the Growth Indicators

To assess the relationship between economic growth and both stock market and banking development, we need: (1) empirical indicators of stock market liquidity, size, volatility, and integration with world capital markets; (2) a measure of banking development; and (3) measures of economic growth and its components. This section first defines six stock market development indicators: one measure of stock market size, two measures of stock market liquidity, a measure of stock market volatility, and two measures of stock market integration with world capital markets. Although each of these indicators has shortcomings, using a variety of measures provides a richer picture of the ties between stock market development and economic growth than if we used only a single indicator. Second, we describe the empirical indicator of banking development. The third subsection defines the growth indicators: real per capita GDP growth, real per capita physical capital stock growth, productivity growth, and the ratio of private savings to GDP. Finally, we present summary statistics on these variables. The Appendix lists data sources, sample periods, and countries.

A. *Stock Market Development Indicators*

1. *Size—Capitalization* measures the size of the stock market and equals the value of listed domestic shares on domestic exchanges divided by GDP. Although large markets do not necessarily function effectively and taxes may distort incentives to list on the exchange, many observers use Capitalization as an indicator of market development.

2. *Liquidity indicators*—We use two related measures of market liquidity. First, *Turnover* equals the value of the trades of domestic shares on domestic exchanges divided by the value of listed domestic shares. Turnover measures the volume of domestic equities traded on domestic exchanges relative to the size of the market. High Turnover is often used as an indicator of low transactions costs. Importantly, a large stock market is not necessarily a liquid market: a large but inactive market will have large Capitalization but small Turnover.

The second measure of market liquidity, *Value Traded,* which equals the value of the trades of domestic shares on domestic exchanges divided by GDP. While not a direct measure of trading costs or the uncertainty associated with trading on a particular exchange, theoretical models of stock market liquidity and economic growth directly motivate Value Traded (Levine, 1991; Bencivenga et al., 1995). Value Traded measures trading volume as a share of national output and should therefore positively reflect liquidity on an economywide basis. Value Traded may be importantly different from Turnover as shown by Demirgüç-Kunt and Levine (1996). While Value Traded captures trading relative to the size of the economy, Turnover measures trading relative to the size of the stock market. Thus, a small, liquid market will have high Turnover but small Value Traded.

Since financial markets are forward looking, Value Traded has one potential pitfall. If markets anticipate large corporate profits, stock prices will rise today. This price rise would increase the value of stock transactions and therefore raise Value Traded. Problematically, the liquidity indicator would rise without a rise in the number of transactions or a fall in transaction costs. This price effect plagues Capitalization too. One way to gauge the influence of the price effect is to look at Capitalization and Value Traded together. The price effect influences both indicators, but only Value Traded is directly related to trading. Therefore, we include both Capitalization and Value Traded indicators together in our regressions. If Value Traded remains significantly correlated with growth while controlling for Capitalization, then the price effect is not dominating the relationship between Value Traded and growth. A second way to gauge the importance of the price effect is to examine Turnover. The price effect does not influence Turnover because stock prices enter the numerator and denominator of Turnover. If Turnover is posi-

and robustly associated with economic growth, then this implies that the price effect is not dominating the relationship between liquidity and long-run economic growth.

3. *International integration measures*—Besides liquidity and size, we use two indicators of the degree of integration with world financial markets to provide evidence on theories that link market integration with economic growth. In perfectly integrated markets, capital flows across international borders to equate the price of risk. If capital controls or other barriers impede capital movements, then the price of risk may differ internationally. To compute measures of integration, we use the international capital asset pricing model (CAPM) and international arbitrage pricing theory (APT).

Since these models are well known, we only cursorily outline the estimation procedures. Both asset pricing models imply that the expected return on each asset is linearly related to a benchmark portfolio or linear combination of a group of benchmark portfolios. Following Robert A. Korajczyk and Claude J. Viallet (1989 p. 562–64), let **P** denote the vector of excess returns on a benchmark portfolio. For the CAPM, **P** is the excess return on a value-weighted portfolio of common stocks. For the APT, **P** represents the estimated common factors based on the excess returns of an international portfolio of assets using the asymptotic principal components technique of Gregory Connor and Korajczyk (1986). Firm-level stock returns from 24 national markets are used to form the value-weighted portfolio for the CAPM and to estimate the common factors for the APT. Given m assets and T periods, consider the following regression:

$$(1) \qquad R_{i,t} = \alpha_i + b_i \mathbf{P}_t + \varepsilon_{i,t},$$

$$i = 1, 2, \ldots, m; \qquad t = 1, 2, \ldots, T,$$

where $R_{i,t}$ is the excess return on asset i in period t, i.e., the return above the return on a risk-free asset or zero-beta asset (an asset with zero correlation with the benchmark portfolio). The $R_{i,t}$'s are based on monthly, firm-level stock returns that have been adjusted for dividends and stock splits. For an average month, there are 551 firms with return data from the 24 markets.

If stock markets are perfectly integrated, then the intercept in a regression of any asset's excess return on the appropriate benchmark portfolio, **P**, should be zero:

$$(2) \qquad \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0.$$

Rejection of the restrictions defined by (2) may be interpreted as rejection of the underlying asset pricing model or rejection of market integration.

Under the assumption that the CAPM and APT are reasonable models of asset pricing, we interpret the monthly estimates of the absolute value of the intercept terms from the multivariate regression (1) as measures of market integration. To compute monthly estimates of stock market integration for each national market, we compute the average of the absolute value of α_i across all stocks in each country each month. Then, we multiply this final value by negative one. Thus, these *CAPM Integration* and *APT Integration* measures are designed to be positively correlated with integration. Moreover, Korajczyk (1996) shows that international integration measures will be negatively correlated with higher official barriers and taxes to international asset trading, bigger transaction costs, and larger impediments to the flow of information about firms.[3]

4. *Volatility*—We measure the volatility of stock returns, *Volatility*, as a 12-month rolling standard deviation estimate that is based on market returns. We cleanse the return series of monthly means and 12 months of autocorrelations using the procedure defined by G. William Schwert (1989). Specifically, we estimate a 12th-order autoregression of monthly returns, R_t, including dummy variables, D_{jt}, to allow for different monthly mean returns:

$$(3) \qquad R_t = \sum_{j=1}^{12} a_j D_{jt} + \sum_{k=1}^{12} b_k R_{t-k} + \nu_t.$$

---

[3] The CAPM and APT Integration measures rely on asset pricing models that the data frequently rejected as good representations of the pricing of risk. For this paper, however, we seek a numerical index of, for example, how much more the United States is integrated into world capital markets than is Nigeria. We are not concerned with whether the index is based at zero. Thus, even if the integration measures include a constant bias, the CAPM and APT Integration measures still provide information on cross-country differences in market integration.

We collect the absolute value of the residuals from equation (3), and then estimate a 12th-order autoregression of the absolute value of the residuals including dummy variables for each month to allow for different monthly standard deviations of returns:

$$(4) \quad |\hat{\nu}| = \sum_{j=1}^{12} c_j D_{jt} + \sum_{k=1}^{12} d_k |\hat{\nu}_{t-k}| + \mu_t .$$

The fitted values from this last equation give estimates of the conditional standard deviation of returns.[4] We include this measure because of the intense interest in market volatility by academics, practitioners, and policy makers.

B. *Banking Development*

An extensive theoretical literature examines the ties between banks and economic activity. Ideally, researchers would construct cross-country measures of how well banks identify profitable activities, exert corporate governance, mobilize resources, manage risk, and facilitate transactions. Economists, however, have not been able to accurately measure these financial services for a broad cross section of countries. Consequently, researchers traditionally use measures of the overall size of the banking sector to proxy for "financial depth" (e.g., Raymond W. Goldsmith, 1969; Ronald I. McKinnon, 1973). Thus, researchers often divide the stock of broad money (M2) by GDP to measure financial depth. As noted by King and Levine (1993a), however, this type of financial depth indicator does not measure whether the liabilities are those of banks, the central bank, or other financial intermediaries, nor does this financial depth measure identify where the financial system allocates capital. Thus, we use the value of loans made by commercial banks and other deposit-taking banks to the private sector divided by GDP, and call this measure *Bank Credit*. Bank Credit improves upon traditional financial depth measures of banking development by isolating credit issued by banks, as opposed to credit issued by the central bank or other intermediaries, and by identifying credit to the private sector, as opposed to credit issued to governments. In our empirical work, we also used traditional measures of financial depth and discuss some of these results below. We focus almost exclusively on the results with Bank Credit.

C. *Channels to Growth*

Besides examining the relationship between these financial development indicators and long-run real per capita GDP growth, *Output Growth*, we also study two channels through which banks and stock markets may be linked to growth: the rate of real per capita physical capital stock growth, *Capital Stock Growth*, and everything else, *Productivity Growth*. Specifically, let Output Growth equal κ*(Capital Stock Growth) + Productivity Growth. To obtain empirical estimates, we (a) obtain Output Growth from national accounts data; (b) use Capital Stock Growth from King and Levine (1994); (c) select a value for κ ($\kappa = 0.3$), and then compute Productivity Growth as a residual.[5] If Capital Stock Growth accurately reflects changes in physical capital and if capacity utilization remains stable when averaged over 18 years, then Productivity Growth should provide a reasonable conglomerate indicator of technological change, quality advances, and resource allocation enhancements.[6]

The last growth indicator we consider, *Savings*, equals gross private savings from Paul Masson et al. (1995). Measuring private saving rates is subject to considerable measurement error, and data on gross private saving are available for many fewer countries in our sample (32) than, for example, Output Growth data (47). Nevertheless, these data offer a unique opportunity to shed some empirical light on important theoretical issues: what is the relationship between private saving rates and stock market liquidity, international risk sharing through integrated capital markets, and the level of banking development?

---

[4] As in Schwert (1989), we use iterated weighted least-squares estimates, iterating three times between (3) and (4), to obtain more efficient estimates.

[5] To compute capital stocks, King and Levine (19 estimate the capital-output ratio for over 100 countries 1950, data permitting, and then iterate forward Robert Summers and Alan Heston (1991) real investment data and a depreciation rate of 0.07. We update these estimates through 1990 using Summers and Heston (19 data. Estimates of the capital share parameter, κ, typi range between 0.25 and 0.40 (see King and Levine [19 for citations). We experimented with values in this and since the results do not importantly change, we the results with $\kappa = 0.3$.

[6] In the regressions, we include a term for inve in human capital.

We term the four variables—Output Growth, Capital Stock Growth, Productivity Growth, and Savings—growth indicators. Thus, this paper evaluates the empirical relationship between the four growth indicators and the six stock market indicators (Turnover, Value Traded, Capitalization, Volatility, CAPM Integration, and APT Integration) plus the banking development indicator (Bank Credit).

D. *Summary Statistics and Correlations*

Table 1 presents summary statistics on the six stock market development indicators, the bank development indicator, and four growth indicators. We have data for a maximum of 47 countries over the 1976–1993 period. Table 1 shows substantial variance among the countries in the growth and financial development indicators. For example, Korea averaged 9.7 percent annual growth over the 1976–1993 period and had a private savings rate of almost 30 percent of GDP, while Cote d'Ivoire grew at −2.5 percent in real per capita terms over the same period and Bangladesh's savings rate was 9 percent of GDP; Taiwan had Value Traded equal to almost 1.2, while Nigeria's Value Traded averaged 0.0002 from 1976–1993.

Table 2 presents correlations. Data permitting, we average the data over the 1976–1993 period so that each country has one observation per variable. We compute the correlations for Capital Stock Growth and Productivity Growth using data averaged over the 1976–1990 period. Three correlations are worth highlighting. First, Bank Credit is highly correlated with the growth indicators and all of the stock market indicators. Second, Bank Credit is very highly correlated with Capitalization (0.65), which suggests that it will be difficult to distinguish between measures of overall size of the equity market and the measure of bank credit to private enterprises divided by GDP. Third, the liquidity measures are positively and significantly correlated with Output Growth, Capital Stock Growth, and Productivity Growth at the 0.05-percent level.

II. Stock Markets, Banks, and Economic Growth

This section evaluates whether measures of banking development and stock market liquidity, size, volatility, and integration with world capital markets are robustly correlated with economic growth, capital accumulation, productivity growth, and private saving rates. The first two subsections use least-squares regressions to study the ties between the growth indicators and measures of banking development, stock market liquidity, market size, and stock return volatility. The next subsection uses instrumental variables to examine the links between the growth indicators, banking development, and measures of capital market integration. We use instrumental variables because the international integration measures are estimated regressors. The final subsection conducts a number of sensitivity checks on the robustness of the results.

A. *Framework: Banking, Liquidity, Size, and Volatility*

This subsection uses cross-country regressions to gauge the strength of the partial correlation between each of the four growth indicators and measures of banking and stock market development. The growth indicators are averaged over the 1976–1993 period. The banking and stock market development indicators are computed at the beginning of the period 1976 (data permitting). There is one observation per country. We organize the investigation around the four stock market development indicators and always control for the level of banking development. Thus, we run 16 basic regressions, where the dependent variable is either Output Growth, Capital Stock Growth, Productivity Growth, or Savings averaged over the 1976–1993 period. The four stock market variables are either Turnover, Value Traded, Capitalization, or Volatility measured at the beginning of the sample period.

TABLE 1—SUMMARY STATISTICS: ANNUAL AVERAGES 1976–1993

	Mean	Median	Maximum	Minimum	Standard deviation	Observa
Output Growth	0.021	0.019	0.097	−0.025	0.022	47
Capital Stock Growth	0.028	0.024	0.095	−0.023	0.026	46
Productivity Growth	0.016	0.014	0.079	−0.019	0.017	46
Savings	20.0	20.8	29.7	9.1	5.1	32
Capitalization	0.32	0.17	2.45	0.01	0.43	46
Value Traded	0.11	0.04	1.16	0.00	0.19	47
Turnover	0.30	0.23	2.05	0.01	0.33	46
Volatility	0.07	0.05	0.31	0.03	0.06	36
Bank Credit	0.80	0.75	2.27	0.12	0.50	47
APT Integration	−4.30	−3.95	−2.17	−6.67	1.48	24
CAPM Integration	−4.08	−3.65	−2.00	−9.98	1.86	24

Notes: Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock grow Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings as a percent of GD Capitalization = value of domestic shares as a share of GDP; Value Traded = value of the trades of domestic shares a share of GDP; Turnover = value of the trades of domestic shares as a share of market capitalization; Volatility = measure of stock return volatility; Bank Credit = bank credit to the private sector as a share of GDP; APT Integra tion = the arbitrage pricing theory measure of stock market integration; CAPM Integration = the international capi asset pricing model measure of stock market integration.

Traditionally, the growth literature uses growth and explanatory variables averaged over long periods. This approach, however, is frequently criticized because: (i) a common shock to the dependent and explanatory variables during the sample period may be driving the empirical findings; and (ii) *contemporaneous regressions*—regressions using dependent and explanatory variables averaged over the same period—do not account for the potential endogenous determination of growth and the explanatory variables. Besides conducting the contemporaneous regressions, we focus on the "initial value" regressions, where we use the values of the banking and stock market indicators in 1976. While this analysis does not resolve the issue of causality, the initial value regressions show that the strong relationship between financial development and the growth indicators does not merely reflect contemporaneous shocks to both, and that stock market and banking d velopment do not simply follow econo development.

To assess the strength of the independ relationship between the initial levels of sto market and banking development and growth variables, we include a wide array control variables, *X*. Specifically, we incl the logarithm of initial real per capital GD *Initial Output,* and the logarithm of the in secondary-school enrollment rate, *Enrollm* because theory and evidence suggest an portant link between long-run growth and tial income and investment in human ca accumulation (Robert M. Solow, 1956; Lu 1988; N. Gregory Mankiw et al., 1992; B and Xavier Sala-i-Martin, 1995). The nu of revolutions and coups, *Revolutions Coups,* is included since many authors that political instability is negatively as ated with economic growth (see Barro

TABLE 2—CORRELATIONS

	Capital Stock Growth	Productivity Growth	Savings	Capitalization	Value Traded	Turnover	CAPM Integration	APT Integration	Volatility	Bank Credit
Output Growth	0.773 (0.001)	0.957 (0.001)	0.4466 (0.008)	0.037 (0.037)	0.522 (0.001)	0.487 (0.001)	0.343 (0.101)	0.28 (0.186)	−0.08 (0.644)	0.347 (0.013)
Capital Stock Growth	—	0.557 (0.001)	0.5300 (0.001)	0.203 (0.171)	0.425 (0.003)	0.356 (0.014)	0.228 (0.296)	0.182 (0.407)	−0.104 (0.547)	0.324 (0.023)
Productivity Growth	—	—	0.4191 (0.014)	0.222 (0.134)	0.417 (0.003)	0.444 (0.002)	0.277 (0.200)	0.209 (0.339)	−0.169 (0.325)	0.372 (0.008)
Savings	—	—	—	−0.0792 (0.656)	0.1601 (0.366)	0.4470 (0.008)	−0.1394 (0.620)	−0.3504 (0.200)	0.1189 (0.555)	0.1189 (0.168)
Capitalization	—	—	—	—	0.331 (0.022)	0.05 (0.735)	0.476 (0.019)	0.36 (0.084)	−0.261 (0.124)	0.647 (0.001)
Value Traded	—	—	—	—	—	0.831 (0.001)	0.188 (0.380)	0.068 (0.752)	0.085 (0.622)	0.449 (0.001)
Turnover	—	—	—	—	—	—	0.074 (0.730)	−0.003 (0.991)	0.186 (0.278)	0.328 (0.023)
CAPM Integration	—	—	—	—	—	—	—	0.78 (0.001)	−0.838 (0.001)	0.45 (0.027)
APT Integration	—	—	—	—	—	—	—	—	0.573 (0.005)	0.454 (0.026)
Volatility	—	—	—	—	—	—	—	—	—	−0.404 (0.014)

p-values in parentheses. Output Growth = real per capital GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Capitalization = value of domestic shares as a share of GDP; Value Traded = value of the trades of domestic shares as a share of GDP; Turnover = value of the trades of domestic shares as a share market capitalization; Volatility = measure of stock return volatility; Bank Credit = bank credit to the private sector as a share of GDP; APT Integration = the arbitrage pricing theory measure of stock market integration; CAPM Integration = the international capital asset pricing model measure of stock market integration.

Sala-i-Martin [1995] for evidence and citations). We also include a variety of macroeconomic indicators in the conditioning information set. The initial values of government consumption expenditures to GDP, *Government*, and the rate of inflation, *Inflation*, are included because theory and some evidence suggests a negative relationship between macroeconomic instability and economic activity (William Easterly and Sergio Rebelo, 1993; Stanley Fischer, 1993; Michael Bruno and Easterly, 1998). Similarly, the initial value of the black market exchange rate premium, *Black Market Premium*, is part of the *X* variables since international price distortions may impede efficient investment decisions and economic growth (David Dollar, 1992). Moreover, the black market premium is a general indicator of policy, price, and trade distortions and therefore is a useful variable to use in assessing the independent relationship between the growth indicators and measures of financial sector development. As discussed below, alternative control variables and combinations of *X* variables do not materially affect the results on the relationship between financial development and economic growth.

B. *Results: Banking, Liquidity, Size, and Volatility*

First, consider the results on stock market liquidity and banking development. Table 3 presents four regressions, where the dependent variable is Output Growth, Capital Stock Growth, Productivity Growth, and Savings,

TABLE 3—INITIAL TURNOVER, BANKS, AND GROWTH, 1976–1993

Independent variables	Dependent variables			
	Output Growth	Capital Stock Growth	Productivity Growth	Savings
Bank Credit	0.0131 (0.0055)	0.0148 (0.0063)	0.0111 (0.0046)	3.8376 (2.3069)
Turnover	0.0269 (0.0090)	0.0222 (0.0094)	0.0201 (0.0088)	7.7643 (5.6864)
R^2	0.5038	0.5075	0.4027	0.4429
Observations	42	41	41	29

Notes: Heteroskedasticity-consistent standard errors in parentheses. Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Bank Credit = initial bank credit to the private sector as a share of GDP; Turnover = initial value of the trades of domestic shares as a share of market capitalization. Other explanatory variables included in each of the regressions: Initial Output, Enrollment, Revolutions and Coups, Government, Inflation, and Black Market Premium.

respectively, and the liquidity measure is initial Turnover. White's heteroskedasticity-consistent standard errors are reported in parentheses. *Both* the stock market liquidity and banking development indicators enter the Output Growth, Capital Stock Growth, and Productivity Growth regressions significantly at the 0.05-percent significance level. To economize on space, we only present the coefficient estimates for the stock market and bank indicators. The full regression results for Table 3 are given in the Appendix [see Table A1]. The other explanatory variables generally enter the regressions as expected. Initial income enters with a significantly negative coefficient and the size of the convergence coefficient is very similar to other studies (Barro and Sala-i-Martin, 1995). Secondary-school enrollment enters the growth regression positively, while political instability enters with a significantly negative coefficient. Although the values of government consumption expenditures divided by GDP and inflation in 1976 enter the growth regression with negative coefficients, they are statistically insignificant, though inflation has a strong negative relationship with capital accumulation and private saving rates. In this sample of countries and with the extensive set of control variables, the black market exchange rate premium does not enter the Output Growth regression significantly, which confirms Levine and Renelt (1992). The growth regression R^2 of 0.50 is consistent with other cross-country growth studies (e.g., Barro and Sala-i-Martin, 1995).

In sum, we find that both the initial level of banking development and the initial level of stock market liquidity have statistically significant relationships with future values of Output Growth, Capital Stock Growth, and Productivity Growth even after controlling for many other factors associated with long-run economic performance. These results are consistent with the view that stock market liquidity and banks facilitate long-run growth (Levine, 1991; Bengt Holmstrom and Jean Tirole, 1993; Bencivenga et al., 1995). The results are not supportive of models that emphasize the negative implications of stock market liquidity (Shleifer and Vishny, 19 Shleifer and Lawrence Summers, 1988).

We do not find a statistically significant link between private saving rates and either stock market liquidity or banking development. Although the saving results should be viewed very skeptically because there are only 29 observations in the regressions, Catherine Bonser-Neal and Kathryn Dewenter (19 find similar results using annual data with 1 observations: there is not a systematic assotion between stock market liquidity and private saving rates. It is also worth noting that these results do not contradict Tullio Jappelli and Marco Pagano's (1994) findings that countries where households are liquidity constrained tend to have higher saving rates. In Jappelli and Pagano (1994), "liquidity constrained" means that households find it relatively difficult to obtain mortgages or consumer credit. In contrast, this paper uses the term liquidity to refer to the ease with which agents can trade equities. Taken together, the two sets of findings imply that countries with large impediments to obtaining mortgage and consumer credit tend to have higher saving rates, while the level of activity on a country's stock exchange is unrelated to saving rates.[7] Furthermore, our finding that stock market liquidity is unrelated to private saving rates is not inconsistent with our finding that stock market liquidity is positively related to physical capital accumulation: (a) Capital Stock Growth is generated by private-sector, public-sector, and foreign investment, while Savings only measures gross private savings of domestic residents; and (b) the savings analysis is based on a much smaller sample of countries.[8] Moreover, while financial development is significantly associated with future Capital Stock Growth, economically, the major channel through which growth is linked to stock markets and banks is through Productivity Growth, not Capital Stock Growth, as we discuss below. Finally, the lack of a strong link between financial-sector development and private savings has implications for Mankiw et al.'s (1992) evaluation of the neoclassical growth model. One weakness in their analysis is that savings rates may be endogenous or proxying for some other country-specific factor. This paper's results suggest that saving rates are not proxying for financial-sector development.

Besides being statistically significant, the estimated coefficients suggest that the relationships between financial-sector development and future rates of long-run growth, capital accumulation, and productivity improvements are economically large. For example, the estimated coefficient implies that a one-standard-deviation increase in initial stock market liquidity (0.3) would increase per capita growth by 0.8 percentage points per year (0.027*0.3) over this period. Accumulating over 18 years, this implies that real GDP per capita would have been over 15 percent higher by 1994 (exp{18*0.008}). The estimated coefficient on Bank Credit also suggests a similarly large economic relationship between banking development and growth. Specifically, a one-standard-deviation increase in initial banking development (0.5) would increase Output Growth by 0.7 percentage points per year (0.013*0.5). Taken together, the results imply that if a county had increased both stock market and banking development in 1976 by one standard deviation, then by 1994 real per capita GDP would have been 31 percent larger, the capital stock per person would have been 29 percent higher, and productivity would have been 24 percent greater. These conceptual experiments do not consider the question of causality nor how to change the financial sector. Nonetheless, the examples illustrate the potentially large economic consequences of stock market liquidity and banking development and the potentially large economic costs of impediments to financial-sector development.

The Value Traded measure of stock market liquidity confirms these findings. Table 4 presents the same type of regressions as in Table 3 except we replace Turnover with Value Traded. Again, the initial liquidity and banking development indicators are significantly and robustly correlated with future rates of economic growth, capital accumulation, and

---

[7] More generally, Jappelli and Pagano (1994 p. 102) note that the finding that financial development is positively linked with economic growth does not contradict their findings, because they focus on "... the effect of imperfections in the mortgage and consumer credit markets, which have no necessary correlation with the development of lending to firms."

[8] It is also true that in the regression analyses, Savings is only available for about 70 percent of the countries for which we have Capital Stock Growth data. However, the [illegible] Neal and Dewenter (1996) findings suggest that the smaller sample is not driving the results. Moreover, we restricted the Capital Stock Growth regressions to the countries with Savings data. While the *t*-statistics on the financial indicators fall, financial development generally remains a significant predictor of Capital Stock Growth even in this smaller sample.

TABLE 4—INITIAL VALUE TRADED, BANKS, AND GROWTH, 1976–1993

Independent variables	Output Growth	Capital Stock Growth	Productivity Growth	Savings
	Dependent variables			
Bank Credit	0.0146 (0.0056)	0.0148 (0.0061)	0.0125 (0.0047)	3.4917 (2.1920)
Value Traded	0.0954 (0.0315)	0.0927 (0.0324)	0.0736 (0.0220)	15.8456 (14.0757)
R^2	0.4655	0.5224	0.3726	0.4278
Observations	43	42	42	29

Notes: Heteroskedasticity-consistent standard errors in parentheses. Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Bank Credit = initial bank credit to the private sector as a share of GDP; Value Traded = initial value of the trades of domestic shares as a share of GDP. Other explanatory variables included in each of the regressions: Initial Output, Enrollment, Revolutions and Coups, Government, Inflation, and Black Market Premium.

productivity growth. Again, the estimated coefficients suggest an economically large relationship between initial financial development and future long-run growth rates. For example, the results imply that if in 1976 Mexico had had the sample mean value of Value Traded (0.046) instead of its actual value of (0.004), annual per capita growth would have been almost 0.4 percentage points faster (0.095*0.04) over the sample period, such that GDP per capita would have been 7.5 percent higher by 1994 (exp{18*0.004}). The economic implications of a symmetric change in banking are even larger. If Mexico had had the sample mean value of banking development in 1976 (0.65) instead of its actual value of (0.13), growth would have been 0.8 percentage points faster per year (0.015*0.52). Combined, these improvements in stock market liquidity and banking development in 1976 are consistent with Mexico enjoying almost 23-percent higher GDP per capita by 1994.

The findings in Tables 3 and 4 also provide some information on the relative importance of the Capital Stock Growth and Productivity Growth channels. For example, the estimated parameter values imply that a one-standard-deviation increase in Value Traded in 1976 (0.2) would increase Output Growth and Capital Stock Growth by about 1.9 percentage points per year. Since growth accounting ercises generally give Productivity Growth weight that is about two times the weight physical capital accumulation (i.e., $\kappa = 1/3$), this implies that Productivity Growth accou for about 1.3 percentage points (1.9 (1/3)*1.9) of the 1.9-percentage-point crease in Output Growth generated by the crease in Value Traded. Thus, the m channel linking financial development w growth runs through Productivity Grow rather than Capital Stock Growth, whi is consistent with the findings in Jo DeGregorio and Pablo E. Guidotti (1995).[9] noted above, the estimated coefficients sho not be viewed as exploitable elasticiti Rather, these conceptual experiments meant to illustrate the economic size of coefficients.

The forward-looking nature of stock pric the "price-effect"—is not driving the str link between market liquidity and the gro indicators. This can be deduced from two sults. First, the price effect does not influe Turnover, and Turnover is robustly lin

---

[9] The Productivity Growth channel is also the main between Bank Credit Output Growth in the Table 3 results.

TABLE 5—INITIAL VALUE TRADED, CAPITALIZATION, BANKS, AND GROWTH, 1976–1993

	Dependent variables			
Independent variables	Output Growth	Capital Stock Growth	Productivity Growth	Savings
Bank Credit	0.0083 (0.0054)	0.0111 (0.0055)	0.0086 (0.0046)	2.9614 (2.0960)
Capitalization	0.0148 (0.0068)	0.0088 (0.0092)	0.0070 (0.0056)	−7.5606 (7.0266)
Value Traded	0.0700 (0.0322)	0.0780 (0.0382)	0.0592 (0.0227)	23.5929 (15.7283)
R^2	0.5186	0.5297	0.4083	0.4499
Observations	42	41	41	29

Notes: Heteroskedasticity-consistent standard errors in parentheses. Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Bank Credit = initial bank credit to the private sector as a share of GDP; Value Traded = initial value of the trades of domestic shares as a share of GDP; Capitalization = initial value of domestic shares as a share of GDP. Other explanatory variables included in each of the regressions: Initial Output, Enrollment, Revolutions and Coups, Government, Inflation, and Black Market Premium.

with future rates of economic growth, capital accumulation, and productivity growth. Second, we include Capitalization and Value Traded together in the same regression to test whether the price-effect is producing the strong empirical links between Value Traded and the growth indicators. The price-effect influences both Capitalization and Value Traded. If the price-effect is driving the empirical association between Value Traded and the growth indicators reported in Table 4, then Value Traded should not remain significantly correlated with the growth indicators when we simultaneously include Capitalization and Value Traded. This is not the case. As reported in Table 5, Value Traded in 1976 remains significantly correlated with future rates of economic growth, capital accumulation, and productivity growth even when controlling for market capitalization (with little change in the estimated coefficients). Thus, the evidence is consistent with the view that expectations of future growth, which are reflected in current stock prices, are driving the strong empirical relationship between stock market liquidity and growth. The evidence is consistent with the view that the ability to trade ownership of an economy's productive technologies easily promotes more efficient resource allocation, capital formation, and faster growth.[10]

Importantly, initial stock market size and stock return volatility are not generally robust predictors of the growth indicators. Although the coefficients presented in Table 6 indicate a positive association between Capitalization and both Output Growth and Capital Stock Growth, this relationship is strongly influenced by a few countries. Specifically, if Jamaica, Korea, and Singapore are removed from the regression, Capitalization no longer

---

[10] The strong link between liquidity and capital accumulation suggests an area for future research. Specifically, three empirical findings need to be reconciled: (1) stock market liquidity is positively tied to capital formation, but (2) equity sales do not finance much of this capital formation (Colin Mayer, 1988), and (3) stock market liquidity is *positively* associated with corporate debt-equity ratios in developing countries (Demirgüç-Kunt and Maksimovic, 1996). These findings imply interactions between stock markets, banks, corporate finance, and corporate investment decisions that many existing theories do not fully capture (though, see Boyd and Smith [1996] and Elisabeth Huybens and Smith [1998]).

TABLE 6—INITIAL CAPITALIZATION, BANKS, AND GROWTH, 1976–1993

	Dependent variables			
Independent variables	Output Growth	Capital Stock Growth	Productivity Growth	Savings
Bank Credit	0.0089	0.0090	0.0094	5.1226
	(0.0061)	(0.0078)	(0.0050)	(2.0927)
Capitalization	0.0230	0.0207	0.0135	−0.7291
	(0.0065)	(0.0081)	(0.0055)	(7.1411)
R^2	0.4577	0.3754	0.3423	0.3189
Observations	45	44	44	31

Notes: Heteroskedasticity-consistent standard errors in parentheses. Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Bank Credit = initial bank credit to the private sector as a share of GDP; Capitalization = initial value of domestic shares as a share of GDP. Other explanatory variables included in each of the regressions: Initial Output, Enrollment, Revolutions and Coups, Government, Inflation, and Black Market Premium.

enters the regression significantly.[11] Similarly, the results on market volatility do not suggest a reliable link to the growth indicators. As shown in Table 7, stock return volatility is not closely linked with future growth, productivity improvements, or private saving rates, and Volatility is positively correlated with capital accumulation. As discussed below, the results on market liquidity are much more robust to the removal of outliers. More importantly, the relationship between stock market size and the growth indicators vanishes when controlling for stock market liquidity (Table 5). Thus, it is not just listing securities on an exchange; it is the ability to trade those securities that is closely tied to economic performance.

C. *International Capital Market Integration, Banking, and the Growth Indicators*

To investigate the relationship between the growth indicators and international capital market integration, we slightly revise the analytical framework in two ways. First, we only have data on capital market integration for 24 countries. Thus, we use pooled cross-section time-series data averaged over the periods 1976–1985 and 1986–1993, so that each country potentially two observations for a maximum 48 observations.[12] Second, CAPM Integratio and APT Integration are estimated regresso Therefore, we use two-stage least squares to d rive consistent standard errors as suggested Adrian Pagan (1984).[13]

Tables 8 and 9 report the results on cap market integration. The CAPM and APT tegration measures enter the growth equatio with a positive coefficient suggesting th greater capital market integration is positiv related to economic performance. Furth more, the point estimates imply a potenti large effect. For example, a one-stand

---

[11] That is, the p-value on the coefficient on Capitalization rises above 0.10.

[12] We choose this asymmetric dividing point b the data for some countries start in 1978.

[13] For instruments, we use Initial Output, Enroll Revolutions and Coups, initial Capitalization, initial Traded, initial Turnover, initial Inflation, initial r international trade to GDP (*Trade*), initial Gover and initial Black Market Premium. The first-stage R^2 0.73 for the CAPM Integration measure and 0.52 f APT Integration measure and the F-statistic for b jects the null hypothesis that none of the cross-se variation in capital market integration is explained explanatory variables. Furthermore, the simple OL gressions yield virtually identical results to the ins tal variable results presented in Tables 8 and 9.

TABLE 7—INITIAL VOLATILITY, BANKS, AND GROWTH, 1976–1993

Independent variables	Dependent variables: Output Growth	Capital Stock Growth	Productivity Growth	Savings
Bank Credit	0.0150 (0.0074)	0.0140 (0.0085)	0.0130 (0.0066)	3.5945 (1.9631)
Volatility	0.0150 (0.0074)	0.4998 (0.1580)	0.0211 (0.2146)	115.0991 (99.4063)
R^2	0.4183	0.6817	0.2938	0.7708
Observations	32	32	32	23

Notes: Heteroskedasticity-consistent standard errors in parentheses. Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Bank Credit = initial bank credit to the private sector as a share of GDP; Volatility = initial measure of stock return volatility. Other explanatory variables included in each of the regressions: Initial Output, Enrollment, Revolutions and Coups, Government, Inflation, and Black Market Premium.

deviation increase in CAPM Integration (1.86) would increase Output Growth by about 1.2 percentage points per year (1.86*0.0065). Nonetheless, the data do not suggest a statistically strong link between capital market integration and the growth indicators. The CAPM and APT Integration measures are not significantly correlated with Output Growth at the 0.10 level. Moreover, the reported regressions exclude Inflation, which is very highly correlated with stock market integration. With inflation included, the *t*-statistics on CAPM Integration and APT Integration become even smaller. While the very small sample may lower confidence in these results, the findings do not support the hypothesis that greater risk sharing through internationally integrated markets affect growth, capital accumulation, productivity growth, or private saving rates.

D. *Sensitivity Analyses*

We conducted a wide array of sensitivity analyses to check the robustness of these results.[14] As mentioned above, regressions using values of the dependent and explanatory variables averaged over the entire sample period yield similar results. Furthermore, changing the conditioning information set did not materially affect our results.[15] For example, altering the set of explanatory variables included in the regression, adding measures of legal efficiency or institutional development, as defined in Paulo Mauro (1995), or using the King and Levine (1993a) measure of financial depth did not affect the strong link between stock market liquidity and growth.[16] We also experimented with an alternative measure of stock market liquidity that gauges trading relative to stock price movements. Specifically, we divide Value Traded by Volatility. All things equal, more liquid markets should be able to support more trading with less price volatility. This alternative measure produced similar results.

We test for the potential influence of outliers in two ways. First, we use the procedure for

---

[14] Unpublished appendices with numerous additional sensitivity analyses are available at http://www.worldbank.org/html/prdmg/grthweb/growth_t.htm.

[15] Furthermore, we used Summers and Heston (1993) data, instead of own currency prices, to compute Government and Output Growth. This did not affect the results.

[16] When the legal efficiency and institutional development indicators are included with enough additional explanatory variables, the sample size falls dramatically, such that the Bank Credit becomes insignificant at the 0.05-percent level in some specifications.

TABLE 8—STOCK MARKET INTEGRATION (CAPM), BANKS, AND GROWTH, 1976–1993, POOLED, INSTRUMENTAL VARIABLES

Independent variables	Dependent variables			
	Output Growth	Capital Stock Growth	Productivity Growth	Savings
Bank Credit	0.0096	0.0143	0.0032	−4.3598
	(0.0134)	(0.0172)	(0.0136)	(2.9495)
CAPM Integration	0.0065	0.0014	0.0085	2.0167
	(0.0043)	(0.0045)	(0.0048)	(2.0609)
Observations	38	38	38	25

Notes: First-stage R^2 for CAPM Integration: 0.73. Heteroskedasticity-consistent standard errors in parentheses. Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Bank Credit = initial bank credit to the private sector as a share of GDP; CAPM Integration = the international capital asset pricing model measure of stock market integration. Instruments: a constant, Initial Output, Enrollment, Revolutions and Coups, and initial values of Government, Black Market Premium, Trade, Capitalization, Value Traded, Turnover, and Bank Credit.

TABLE 9—STOCK MARKET INTEGRATION (APT), BANKS, AND GROWTH, 1976–1993, POOLED, INSTRUMENTAL VARIABLES

Independent variables	Dependent variables			
	Output Growth	Capital Stock Growth	Productivity Growth	Savings
Bank Credit	0.0148	0.0186	0.0117	−3.8182
	(0.0143)	(0.0166)	(0.0150)	(2.3952)
APT Integration	0.0075	−0.0008	0.0086	2.8466
	(0.0074)	(0.0076)	(0.0073)	(1.7108)
Observations	38	38	38	25

Notes: First-stage R^2 for APT Integration: 0.52. Heteroskedasticity-consistent standard errors in parentheses. Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Bank Credit = initial bank credit to the private sector as a share of GDP; APT Integration = the arbitrage pricing theory measure of stock market integration. Instruments: a constant, Initial Output, Enrollment, Revolutions and Coups, and initial values of Government, Black Market Premium, Trade, Capitalization, Value Traded, Turnover, and Bank Credit.

analyzing the influence of particular observations described in William Greene (1993 pp. 287–88). This procedure identifies countries that exert a large effect on each equation's residuals. Using a critical value of 2.5, we find that removing particularly influential observations does not affect our conclusions. Second, we use a more subjective method for identifying influential observations; we use scatterplots of the partial relationship between each of the growth indicators and the individual stock market cators to identify outliers that may be sively influencing the slope and significan the estimated regression line.[17] Removing

---

[17] Specifically, in the multivariate regression of on X, Bank Credit, and $S(k)$, where $S(k)$ repres particular stock market indicator taken in turn, the scatterplot is computed as follows: regress $G(i)$

TABLE 10—INITIAL STOCK MARKET DEVELOPMENT, BANKS, AND GROWTH, 78-COUNTRY SAMPLE

Dependent variable: Output Growth	78-country sample		Original sample	
Stock market indicator (SMI):	Bank Credit	SMI	Bank Credit	SMI
Turnover	0.015 (2.753)	0.022 (2.448)	0.013 (2.753)	0.027 (2.448)
Value Traded	0.013 (2.630)	0.111 (4.242)	0.015 (2.536)	0.095 (3.854)
Capitalization	0.012 (2.047)	0.018 (2.826)	0.009 (1.866)	0.023 (2.672)

Notes: Heteroskedasticity-consistent *t*-statistics in parentheses. Output Growth = real per capita GDP growth; Bank Credit = initial bank credit to the private sector as a share of GDP; Turnover = initial value of trades of domestic shares as a share of market capitalization; Capitalization = initial value of domestic shares as a share of GDP; Value Traded = initial value of trades of domestic shares as a share of GDP; Other explanatory variables included in each of the regressions: Initial Output, Enrollment, Revolutions and Coups, Government, Inflation, and Black Market Premium.

ential observations importantly weakens the relationship between the growth indicators and market size, as noted above. The other results do not change. In particular, stock market liquidity remains robustly correlated with growth, capital accumulation, and productivity growth after removing potential outliers.

We were also concerned about a potential sample selection problem: we only include countries with sufficient stock market activity to warrant inclusion in data bases. We have data on all the non-stock market data for an additional 31 countries. Although we do not have explicit observations on stock transactions in these economies, anecdotal information and a review of official documents suggest that stock market activity in these countries was inconsequential in 1976. Thus, for these 31 countries, we enter values of zero for Capitalization, Value Traded, and Turnover.[18] Zero is not an extreme guess. Recall from Table 1 that the minimum values for Capitalization, Value Traded, and Turnover are 0.01, 0.0002, and 0.006 with standard deviations of 0.43, 0.19, and 0.33, respectively. As shown in Table 10, the link between economic growth and the initial levels of both stock market liquidity and banking development remains strong even when including data on these additional 31 countries.[19]

III. Conclusion

This paper studied the empirical relationship between various measures of stock market development, banking development, and long-run economic growth. We find that,

---

Bank Credit and collect the residuals, $U(G(i))$. Regress $S(k)$ on X and Bank Credit and collect the residuals, $U(S(k))$. Then plot $U(G(i))$ against $U(S(k))$. This gives a two-dimensional graph of the relationship between $G(i)$ and $S(k)$ controlling for X and Bank Credit. This helps identify particularly influential observations.

[18] These 31 countries are Bolivia, Botswana, Cameroon, Central African Republic, Costa Rica, Dominican Republic, Ecuador, Ethiopia, Ghana, Guatemala, Guyana, Haiti, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Nicaragua, Niger, Paraguay, Rwanda, Senegal, Somalia, Sri Lanka, Tunisia, Uruguay, Zaire, and Zambia.

[19] Using these additional 31 countries does not alter the conclusions about the robust links between the financial indicators and Capital Stock Growth and Productivity Growth.

even after controlling for many factors associated with growth, stock market liquidity and banking development are both positively and robustly correlated with contemporaneous and future rates of economic growth, capital accumulation, and productivity growth. This result is consistent with the view that a greater ability to trade ownership of an economy's productive technologies facilitates efficient resource allocation, physical capital formation, and faster economic growth. Furthermore, since measures of stock market liquidity and banking development both enter the growth regressions significantly, the findings suggest that banks provided different financial services from those provided by stock markets. Thus, to understand the relationship between the financial system and long-run growth more comprehensively, we need theories in which both stock markets and banks arise and develop simultaneously while providing different bundles of financial services to the economy. We find no support for the contentions that stock market liquidity, international capital market integration, or stock return volatility reduce private saving rates or hinder long-run growth. This paper finds a strong, positive link between financial development and economic growth and the results suggest that financial factors are an integral part of the growth process.

DATA APPENDIX

A. *Variables and Sources*

Data are available at the web site http://www.worldbank.org/html/prdmg/grthweb/growth_t.htm.

CAPM Integration and *APT Integration:* Measure of each stock market's integration with world equity markets based on the capital asset pricing model and arbitrage pricing theory, respectively. (Sources: Korajczyk, 1994, 1996.)

Bank Credit: Stock of credit by commercial and deposit-taking banks to the private sector divided by GDP. (Source: International Monetary Fund's (IMF's) *International Financial Statistics.*)

Black Market Premium: Black market exchange rate premium. (Sources: *Picks Currency Yearbook* through 1989 and *World Currency Yearbook.*)

Capital Stock Growth: Growth rate in capital stock per person, available through 1990. (Sources: King and Levine, 1994.)

Capitalization: Average value of listed domestic shares on domestic exchanges in a year divided by GDP that year. (Sources: International Finance Corporation's (IFC's) *Emerging Markets Data Base* (electronic version) and the IMF's *International Financial Statistics.*)

Government: Government consumption share of GDP. (Sources: IMF's *International Financial Statistics* and World Bank's *World Development Indicators.*)

Inflation: Rate of change in the GDP deflator; if unavailable, consumer price index is used. (Sources: IMF's *International Financial Statistics* and World Bank's *World Development Indicators.*)

Initial Output: Logarithm of real per capita GDP in 1976. (Source: IMF's *International Financial Statistics.*)

Enrollment: Logarithm of the secondary school enrollment rate in 1976. (Sources: IMF's *International Financial Statistics* and World Bank's *World Development Indicators.*)

Output Growth: Growth of real per capita gross domestic product. (Source: IMF's *International Financial Statistics.*)

Productivity Growth: Output Growth minus 0.3 times Capital Stock Growth, available through 1990. (Source: King and Levine, 1994.)

Revolutions and Coups: Number of revolutions and coups per year, averaged over the 1980's. (Source: Arthur S. Banks, 1994.)

Savings: Gross private saving as a percent of GDP, available from 1982 onward for countries classified as "developing" by the IMF and for the entire sample period for industrial countries. (Source: Masson et al., 1995.)

Trade: Exports plus imports divided by GDP. (Sources: IMF's *International Financial Statistics* and World Bank's *World Development Indicators.*)

Turnover: Value of the trades of domestic shares on domestic exchanges over the year divided by the average value of domestic shares listed on domestic exchanges in year. (Sources: IFC's *Emerging Markets Data Base* (electronic version) and the IMF's *International Financial Statistics.*)

Value Traded: Value of the trades of domestic shares on domestic exchanges over the year divided by GDP. (Sources: IFC's *Emerging Markets Data Base* (electronic version) and the IMF's *International Financial Statistics.*)

Volatility: Measure of the volatility of stock returns, based on the stock market index value. (Sources: IFC's *Emerging Markets Data Base* (electronic version) and the IMF's *International Financial Statistics.*)

B. *Countries Coverage and Sample Period*

The following countries were used in the analyses: Argentina (i, v), Australia (i, s, v), Austria (s, v), Bangladesh (s), Belgium (i, v), Brazil (i, v), Canada (s, v), Chile (i, s, v), Colombia (i, s, v), Cote d'Ivoire, Germany (s, v), Denmark (s, v), Egypt (s), Spain (s, v), Finland (s, v), France (s, v), United Kingdom (i, s, v), Greece (i, s, v), Hong Kong, Indonesia (i, s), India (i, s, v), Israel (v), Italy (i, s, v), Jamaica (s), Jordan (i, v), Japan (i, s, v), Korea (i, s, v), Luxembourg, Mexico (i, v), Malaysia (i, s, v), Morocco (s), Nigeria (i, s), The Netherlands (s, v), Norway (s, v), New Zealand (s, v), Pakistan (i, v), Peru, Philippines (i, v), Portugal (i, s, v), Singapore, Sweden (s, v), Thailand (i, v), Turkey (s, v), Taiwan (i, v), United States (i, s, v), Venezuela (i, v), and Zimbabwe (i, s, v).

The "v" in parentheses indicates that this country is one of the 36 countries for which we computed Volatility from monthly stock returns. The "i" in parentheses indicates that this country is one of the 24 with CAPM and APT Integration data in Korajczyk (1994, 1996). The "s" in parentheses indicates that this country is one of the 32 countries with private savings data in Masson et al. (1995). Unless indicated otherwise, the data are averages over the period 1976–1993.

Table A1 follows.

TABLE A1—COMPLETE TABLE 3 RESULTS—INITIAL TURNOVER, BANKS, AND GROWTH, 1976–1993

	Dependent variables			
Independent variables	Output Growth	Capital Stock Growth	Productivity Growth	Savings
Constant	0.0464	0.1049	0.0324	29.2948
	(0.0246)	(0.0341)	(0.0150)	(6.0756)
Initial Output	−0.0139	−0.0120	−0.0078	−0.5831
	(0.0049)	(0.0073)	(0.0042)	(1.8875)
Enrollment	0.0230	0.0049	0.0118	−0.3602
	(0.0125)	(0.0152)	(0.0097)	(4.7179)
Revolutions and Coups	−0.0346	−0.0306	−0.0227	−13.0141
	(0.0108)	(0.0113)	(0.0083)	(4.3871)
Government	−0.0619	−0.0021	−0.0407	−21.5703
	(0.0379)	(0.0532)	(0.0031)	(20.6724)
Inflation	−0.0071	−0.0296	−0.0085	−11.3403
	(0.0065)	(0.0107)	(0.0082)	(5.9731)
Black Market Premium	0.000	−0.0002	0.0000	−0.0036
	(0.0000)	(0.0001)	(0.0000)	(0.0204)
Bank Credit	0.0131	0.0148	0.0111	3.8376
	(0.0055)	(0.0063)	(0.0046)	(2.3069)

TABLE A1—*Continued.*

	Dependent variables		
Independent variables	Output Growth	Capital Stock Growth	Productivity Growth
Turnover	0.0269 (0.0090)	0.0222 (0.0094)	0.0201 (0.0088)
R^2	0.5038	0.5075	0.4027
Observations	42	41	41

Notes: Heteroskedasticity-consistent standard errors in parentheses. Output Growth = real per capita GDP growth; Capital Stock Growth = real per capita capital stock growth; Productivity Growth = Output Growth-(0.3) (Capital Stock Growth); Savings = private savings divided by GDP; Initial Output = logarithm of initial real per capita GDP; Enrollment = logarithm of initial secondary school enrollment; Revolutions and Coups = number of revolutions and coups per year; Government = initial government consumption expenditures divided by GDP; Inflation = initial inflation rate; Black Market Premium = initial black market exchange rate premium; Bank Credit = initial bank credit to the private sector as a share of GDP; Turnover = initial value of the trades of domestic shares as a share of market capitalization.

REFERENCES

Atje, Raymond and Jovanovic, Boyan. "Stock Markets and Development." *European Economic Review*, April 1993, *37*(2/3), pp. 632–40.

Bagehot, Walter. *Lombard Street*. Homewood, IL: Irwin, 1873.

Banks, Arthur S. "Cross-National Time Series Data Archive." Center for Social Analysis, State University of New York, Binghamton, 1994.

Barro, Robert J. "Economic Growth in a Cross Section of Countries." *Quarterly Journal of Economics*, May 1991, *56*(2), pp. 407–43.

Barro, Robert J. and Sala-i-Martin, Xavier. *Economic growth*. New York: McGraw-Hill, 1995.

Bencivenga, Valerie R. and Smith, Bruce D. "Financial Intermediation and Endogenous Growth." *Review of Economic Studies*, April 1991, *58*(2), pp. 195–209.

Bencivenga, Valerie R.; Smith, Bruce D. and Starr, Ross M. "Transactions Costs, Technological Choice, and Endogenous Growth." *Journal of Economic Theory*, October 1995, *67*(1), pp. 53–177.

Bhide, Amar. "The Hidden Costs of Stock Market Liquidity." *Journal of Financial Economics*, August 1993, *34*(2), pp. 31–51.

Bonser-Neal, Catherine and Dewenter, Kathryn. "Does Financial Market Development Stimulate Savings? Evidence From Emerging Market Stock Markets." Mimeo, University of Indiana, 1996.

Boyd, John H. and Prescott, Edward C. "Financial Intermediary-Coalitions." *Journal of Economics Theory*, April 1986, *38*(2), pp. 211–32.

Boyd, John H. and Smith, Bruce D. "The Coevolution of the Real and Financial Sectors in the Growth Process." *World Bank Economic Review*, May 1996, *10*(2), pp. 371–96.

Bruno, Michael and Easterly, William. "Inflation Crises and Long-Run Growth." *Journal of Monetary Economics*, March 1998, *41*(1), pp. 3–26.

Connor, Gregory and Korajczyk, Robert A. "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis." *Journal of Financial Economics*, March 1986, *15*(3), pp. 373–94.

DeGregorio, Jose and Guidotti, Pablo E. "Financial Development and Economic Growth." *World Development*, March 1995, *23*(3), pp. 433–48.

De Long, J. Bradford; Shleifer, Andrei; Summers, Lawrence H. and Waldmann, Robert J. "The Size and Incidence of the Losses from Noise Trading." *Journal of Finance*, July 1989, *44*(3), pp. 681–96.

Demirgüç-Kunt, Asli and Levine, Ross. "Stock Market Development and Financial Intermediaries: Stylized Facts." *World Bank Economic Review*, May 1996, *19*(2), 291–322.

Demirgüç-Kunt, Asli and Maksimovic, Vojislav. "Financial Constraints, Uses of Funds, and Firm Growth: An International Comparison." Mimeo, World Bank, 1996.

Devereux, Michael B. and Smith, Gregor W. "International Risk Sharing and Economic Growth." *International Economic Review*, August 1994, *35*(4), pp. 535–50.

Diamond, Douglas W. "Financial Intermediation and Delegated Monitoring." *Review of Economic Studies*, July 1984, *51*(3), pp. 393–414.

Dollar, David. "Outward-Oriented Developing Economies Really Do Grow More Rapidly: Evidence from 95 LDCs, 1976–1985." *Economic Development and Cultural Change*, April 1992, *40*(3), pp. 523–44.

Easterly, William and Rebelo, Sergio. "Fiscal Policy and Economic Growth: An Empirical Investigation." *Journal of Monetary Economics*, December 1993, *32*(3), pp. 417–58.

Fischer, Stanley. "The Role of Macroeconomic Factors in Growth." *Journal of Monetary Economics*, December 1993, *32*(3), pp. 485–511.

Goldsmith, Raymond W. *Financial structure and development*. New Haven, CT: Yale University Press, 1969.

Greene, William H. *Economic analysis*. Englewood Cliffs, NJ: Prentice Hall, 1993.

Greenwood, Jeremy and Jovanovic, Boyan. "Financial Development, Growth, and the Distribution of Income." *Journal of Political Economy*, October 1990, Pt. 1, *98*(5), pp. 1076–107.

Holmstrom, Bengt and Tirole, Jean. "Market Liquidity and Performance Monitoring." *Journal of Political Economy*, August 1993, *101*(4), pp. 678–709.

Huybens, Elisabeth and Smith, Bruce D. "Inflation, Financial Markets, and Long-Run Growth." *Journal of Monetary Economics*, 1998 (forthcoming).

International Finance Corporation. *Emerging markets data base*. Washington, DC: International Finance Corporation, various issues.

International Monetary Fund. *International financial statistics*. Washington, DC: International Monetary Fund, various issues.

Jappelli, Tullio and Pagano, Marco. "Saving, Growth, and Liquidity Constraints." *Quarterly Journal of Economics*, February 1994, *109*(1), pp. 93–109.

King, Robert G. and Levine, Ross. "Finance and Growth: Schumpeter Might Be Right." *Quarterly Journal of Economics*, August 1993a, *108*(3), pp. 717–38.

______. "Finance, Entrepreneurship, and Growth: Theory and Evidence." *Journal of Monetary Economics*, December 1993b, *32*(3), pp. 513–42.

______. "Capital Fundamentalism, Economic Development, and Economic Growth." *Carnegie-Rochester Series on Public Policy*, June 1994, *40*, pp. 259–92.

Korajczyk, Robert A. "Measuring Integration of Developed and Emerging Markets." Mimeo, Northwestern University, 1994.

______. "A Measure of Stock Market Integration for Developed and Emerging Markets." *World Bank Economic Review*, May 1996, *10*(2), pp. 267–89.

Korajczyk, Robert A. and Viallet, Claude J. "An Empirical Investigation of International Asset Pricing." *Review of Financial Studies*, September 1989, *2*(4), pp. 553–85.

Levine, Ross. "Stock Markets, Growth, and Tax Policy." *Journal of Finance*, September 1991, *46*(4), pp. 1445–65.

______. "Financial Development and Economic Growth: Views and Agenda." *Journal of Economic Literature*, June 1997, *35*(2), pp. 688–726.

Levine, Ross and Renelt, David. "A Sensitivity Analysis of Cross-Country Growth Regressions." *American Economic Review*, September 1992, *82*(4), pp. 942–63.

Lucas, Robert E., Jr. "On the Mechanics of Economic Development." *Journal of Monetary Economics*, July 1988, 22(1), pp. 3–42.

Mankiw, N. Gregory; Romer, David and Weil, David. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics*, May 1992, *107*(2), pp. 407–37.

Masson, Paul; Bayoumi, Tamim and Samiei, Hossein. "Saving Behavior in Industrial and Developing Countries." International Monetary Fund Staff Studies for the

World Economic Outlook, September 1995.

Mauro, Paulo. "Corruption and Growth." *Quarterly Journal of Economics*, August 1995, *110*(3), pp. 681–712.

Mayer, Colin. "New Issues in Corporate Finance." *European Economic Review*, June 1988, *32*(5), pp. 1167–88.

McKinnon, Ronald I. *Money and capital in economic development*. Washington, DC: Brookings Institution, 1973.

Obstfeld, Maurice. "Risk-Taking, Global Diversification, and Growth." *American Economic Review*, December 1994, *84*(5), pp. 1310–29.

Pagan, Adrian. "Econometric Issues in the Analysis of Regressions with Generated Regressors." *International Economic Review*, February 1984, *25*(1), pp. 221–47.

Picks Currency Yearbook. New York: Picks, various issues.

Rajan, Raghuram G. and Zingales, Luigi. "Financial Dependence and Growth." *American Economic Review*, June 1998, *88*(3), pp. 559–86.

Robinson, Joan. "The Generalization of the General Theory." *The rate of interest and other essays*. London: Macmillan, 1952, pp. 67–146.

Schumpeter, Joseph A. *Theorie der wirtschaftlichen entwicklung*. Leipzig, Germany: Dunker & Humblot, 1912.

Schwert, G. William. "Why Does Stock Market Volatility Change Over Time?" *Journal of Finance*, December 1989, *49*(5), pp. 1115–53.

Shleifer, Andrei and Summers, Lawrence. "Breach of Trust in Hostile Takeovers," in A. Auerbach, ed., *Corporate takeovers: Causes and consequences*. Chicago: University of Chicago Press, 1988, pp. 33–56.

Shleifer, Andrei and Vishy, Robert W. "Large Shareholders and Corporate Control." *Journal of Political Economy*, June 1986, *96*(3), pp. 461–88.

Solow, Robert M. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics*, February 1956, *70*(1), pp. 65–94.

Summers, Robert and Heston, Allan. "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950–1988." *Quarterly Journal of Economics*, May 1991, *106*(2), pp. 327–68.

———. "Penn World Tables, Version 5.5," available on diskette from the National Bureau of Economic Research, Cambridge, MA, 1993.

Williamson, Stephen D. "Costly Monitoring, Financial Intermediation, and Equilibrium Credit Rationing." *Journal of Monetary Economics*, September 1986, *18*(2), pp. 159–79.

World Bank. *World development indicators*. Washington, DC: World Bank, various issues.

World Currency Yearbook. New York: International Currency Analysis, Inc., various issues.

Financial Dependence and Growth

By RAGHURAM G. RAJAN AND LUIGI ZINGALES [*]

This paper examines whether financial development facilitates economic growth by scrutinizing one rationale for such a relationship: that financial development reduces the costs of external finance to firms. Specifically, we ask whether industrial sectors that are relatively more in need of external finance develop disproportionately faster in countries with more-developed financial markets. We find this to be true in a large sample of countries over the 1980's. We show this result is unlikely to be driven by omitted variables, outliers, or reverse causality. (*JEL* O4, F3, G1)

A large literature, dating at least as far back as Joseph A. Schumpeter (1911), emphasizes the positive influence of the development of a country's financial sector on the level and the rate of growth of its per capita income. The argument essentially is that the services the financial sector provides—of reallocating capital to the highest value use without substantial risk of loss through moral hazard, adverse selection, or transactions costs—are an essential catalyst of economic growth. Empirical work seems consistent with this argument. For example, on the basis of data from 35 countries between 1860 and 1963, Raymond W. Goldsmith (1969 p. 48) concludes that "a rough parallelism can be observed between economic and financial development if periods of several decades are considered." Nevertheless, studies such as these simply suggest correlation. As Goldsmith puts it: "There is no possibility, however, of establishing with confidence the direction of the causal mechanism, i.e., of deciding whether financial factors were responsible for the acceleration of economic development or whether financial development reflected economic growth whose mainsprings must be sought elsewhere." While Goldsmith is agnostic, other economists have expressed downright skepticism that financial development is anything but a sideshow to economic development. Joan Robinson (1952 p. 86) is representative of such a viewpoint when she claims "where enterprise leads, finance follows."

In an important recent paper, Robert G. King and Ross Levine (1993a) investigate the causality problem following a *post hoc, ergo propter hoc* approach. They show that the predetermined component of financial development is a good predictor of growth over the next 10 to 30 years. However, the skeptic could still offer a number of arguments against attributing causality.

First, both financial development and growth could be driven by a common omitted variable such as the propensity of households in the economy to save. Since endogenous savings (in certain models of growth) affects the long-run growth rate of the economy, it may not be surprising that growth and initial financial development are correlated. This argument is also hard to refute with simple cross-country regressions. In the absence of a well-accepted theory of growth, the list of potential omitted variables that financial-sector development might be a proxy for is large, and the explanatory variables to include a matter of conjecture.

Second, financial development—typically measured by the level of credit and the size of the stock market—may predict economic growth simply because financial markets anticipate future growth; the stock market capitalizes the present value of growth opportunities, while financial institutions lend more if they think sectors will grow. Thus financial development may simply be a leading indicator rather than a causal factor.

---

[*] Graduate School of Business, University of Chicago, 1101 E. 58th St., Chicago, IL 60637. We thank George Benston, Marco Da Rin, Eugene Fama, Peter Klenow, Krishna Kumar, Ross Levine, Jonathan Macy, Colin Mayer, Canice Prendergast, Andres Rodriguez-Clare, David Scharfstein, Robert Vishny, and two anonymous referees for valuable comments. Jayanta Sen, Dmitrii Kachintsev, and Alfred Shang provided excellent research assistance. A preliminary study was supported by the World Bank. We gratefully acknowledge financial support from NSF Grant No. SBR-9423645.

One way to make progress on causality is to focus on the details of theoretical mechanisms through which financial development affects economic growth, and document their working. Our paper is an attempt to do this. Specifically, theorists argue that financial markets and institutions help a firm overcome problems of moral hazard and adverse selection, thus reducing the firm's cost of raising money from outsiders. So financial development should disproportionately help firms (or industries) typically dependent on external finance for their growth. Such a finding could be the "smoking gun" in the debate about causality. There are two virtues to this simple test. First, it looks for evidence of a specific mechanism by which finance affects growth, thus providing a stronger test of causality. Second, it can correct for fixed country (and industry) effects. Though its contribution depends on how reasonable our microeconomic assumptions are, it is less dependent on a specific macroeconomic model of growth.

We construct the test as follows. We identify an industry's need for external finance (the difference between investments and cash generated from operations) from data on U.S. firms. Under the assumption that capital markets in the United States, especially for the large listed firms we analyze, are relatively frictionless, this method allows us to identify an industry's technological demand for external financing. Under the further assumption that such a technological demand carries over to other countries, we examine whether industries that are more dependent on external financing grow relatively faster in countries that, a priori, are more financially developed.

This would imply that, ceteris paribus, an industry such as Drugs and Pharmaceuticals, which requires a lot of external funding, should develop relatively faster than Tobacco, which requires little external finance, in countries that are more financially developed. Consider, for instance, Malaysia, Korea, and Chile, which are moderate-income, fast-growing, countries that differ considerably in their financial development. Consistent with our hypothesis, in Malaysia, which was the most financially developed by our measure, Drugs and Pharmaceuticals grew at a 4-percent higher annual real rate over the 1980's than Tobacco (the growth rate for each industry is adjusted for the worldwide growth rate of that industry). In Korea, which was moderately financially developed, Drugs grew at a 3-percent higher rate than Tobacco. In Chile, which was in the lowest quartile of financial development, Drugs grew at a 2.5-percent *lower* rate than Tobacco. So financial development seems to affect relative growth rates of industries in the way predicted. We establish this result more systematically for a larger cross section of industries and countries in the body of the paper.

Delving deeper into the components of growth, industry growth can be decomposed into the growth in the number of establishments and the growth in the average size of existing establishments. New establishments are more likely to be new firms, which depend more on external finance than established firms. So the growth of the number of establishments in industries dependent on external finance should be particularly sensitive to financial development. This is indeed the case. Our estimates suggest that financial development has almost twice the economic effect on the growth of the number of establishments as it has on the growth of the average size of establishments. This suggests that an additional indirect channel through which financial development could influence growth is by disproportionately improving the prospects of young firms. If these are typically innovators, they make possible Schumpeterian "waves of creative destruction" that would not even get initiated in countries with less-developed markets.

Let us be careful about what we find, and about what we have little to say. Our findings suggest that the *ex ante* development of financial markets facilitates the *ex post* growth of sectors dependent on external finance. This implies that the link between financial development and growth identified elsewhere may stem, at least in part, from a channel identified by the theory: financial markets and institutions reduce the cost of external finance for firms. Of course, our analysis suggests only that financial development liberates firms from the drudgery of generating funds internally. It is ultimately the availability of profitable investment opportunities that drives growth, and we have little to say about where these come from. In the imagery of Rondo Cameron (1967 p. 2), we find evidence consistent with finance as a lubricant, essential no doubt, but not a substitute for the machine.

Our paper relates closely to three recent papers that attempt to establish the direction of causation of the finance-growth correlation. Asli Demirgüç-Kunt and Vojislav Maksimovic (1996) also use micro-data to develop a test of the influence of financial development on growth. Using firm-level data, they estimate the proportion of firms whose rate of growth exceeds the growth that could have been supported only by internal resources. They then run a cross-country regression and find that this proportion is positively related to the stock market turnover and to a measure of law enforcement. There are two essential differences from our paper. First, their estimate of the internal growth rate of a firm is dependent on the firm's characteristics. While it is potentially more accurate than our measure of external dependence, it is also more endogenous. Second, they focus on between-country differences in the spirit of traditional cross-country regressions, while our focus is on within-country, between-industry differences. The latter is an important innovation in this paper.

Jith Jayaratne and Philip E. Strahan (1996) examine the liberalization of the banking sector in different states in the United States in recent years and show that this had a positive influence on a state's growth. Our attempt to correct for fixed effects is similar to theirs. They use differences in growth rates across the temporal shock of liberalization while we use differences between industries within a country to do so. Since they focus on a very nice natural experiment to provide identification, their methodology may be harder to apply to different countries or different questions. But the more important difference is that we focus on providing evidence for a microeconomic channel through which finance is supposed to work rather than examining, as they do, the broader correlation between finance and growth.

Finally, Levine and Sarah Zervos (1998) study whether stock markets and banks promote economic growth. They find that measures of market liquidity are strongly related to growth, capital accumulation, and productivity, while surprisingly, more traditional measures of development such as stock market size are not as robustly correlated. They also find that bank lending to the private sector has a strong independent effect on growth. They focus on a richer set of measures of financial development and growth than we do, but their cross-country regression methodology is also more traditional. The two studies should be viewed as complementary—theirs providing information on a broader set of correlations, while ours details a mechanism.

In this paper, we start by describing the theoretical underpinnings of our work in Section I and then our measure of external dependence in Section II. In Section III, we present our data on financial development, country characteristics, and industry growth. In Section IV we set up our main test and discuss the results. We explore other tests and the robustness of our findings in Section V. Section VI concludes.

I. Theoretical Underpinnings and the Basic Test

A. *Theoretical Underpinnings*

There has been extensive theoretical work on the relationship between financial development and economic growth. Economists have emphasized the role of financial development in better identifying investment opportunities, reducing investment in liquid but unproductive assets, mobilizing savings, boosting technological innovation, and improving risk taking.[1] All these activities can lead to greater economic growth. We do not have the space to go into all these theories [see Levine (1997) for a comprehensive recent survey] so we content ourselves with outlining the essential theoretical underpinnings for our test.

---

[1] Apart from the papers discussed below, see Valerie R. Bencivenga and Bruce D. Smith (1991), Gilles Saint-Paul (1992), King and Levine (1993b), Maurice Obstfeld (1994), and John H. Boyd and Smith (1996).

Jeremy Greenwood and Boyan Jovanovic (1990) develop a model where the extent of financial intermediation and economic growth are endogenously determined. In their model, financial intermediaries can invest more productively than individuals because of their better ability to identify investment opportunities. So financial intermediation promotes growth because it allows a higher rate of return to be earned on capital, and growth in turn provides the means to implement costly financial structures.

Equivalently, the model could be recast to show that financial development reduces the cost of raising funds from sources external to the firm relative to the cost of internally generated cash flows. External funds are generally thought to be costlier because outsiders have less control over the borrower's actions (see, for example, Michael C. Jensen and William R. Meckling, 1976) or because they know less about what the borrower will do with the funds (see Joseph E. Stiglitz and Andrew Weiss, 1981; Stewart C. Myers and Nicholas S. Majluf, 1984). Financial development, in the form of better accounting and disclosure rules, and better corporate governance through institutions, will reduce the wedge between the cost of internal and external funds and enhance growth, especially for firms that are most reliant on external financing.[2]

A second issue is how financial development takes place. Some economists take the development of the financial market as exogenous to the model arguing that "differences in the extent of financial markets across countries seem to depend primarily on legislation and government regulation" (Bencivenga and Smith, 1991 p. 207). By contrast, Greenwood and Jovanovic (1990) have a "once-and-for-all" lump-sum cost of development and development is endogenous to their framework. From the perspective of our paper, it really does not matter whether legal and political or economic forces are responsible for financial development. Our focus is on whether the predetermined level of financial development affects growth. All we need for the stock of financial development to matter even when development is endogenous is that there be a cost to development (as in Greenwood and Jovanovic) or that financial development cannot happen instantaneously (as in reputational models of financial development such as Douglas W. Diamond, 1989). Either assumption seems plausible.

If financial development cannot take place at low cost and on the fly, the above theories would suggest that the a priori existence of a well-developed financial market should disproportionately improve the *ex post* growth rates of industries that are technologically more dependent on external funds.

B. *The Basic Test*

The most disaggregated comprehensive data on growth that we have for countries is at the industry level (data at the firm level, if available, is typically limited to large listed firms). Our hypothesis is that industries that are more dependent on external financing will have relatively higher growth rates in countries that have more developed financial markets.

Therefore, the dependent variable is the average annual real growth rate of value added in industry j in country k over the period 1980–1990. If we can measure industry j's dependence on external finance and country k's financial development, then after correcting for country and industry effects we must find that the coefficient estimate for the interaction between dependence and development is positive.

---

[2] In Greenwood and Jovanovic (1990), there are no moral hazard or asymmetric information problems at the level of the entrepreneur. The intermediary simply provides information about economywide trends that the entrepreneur cannot figure out for himself, enabling the entrepreneur to invest his own funds more productively. An equivalent formulation is to distinguish between savers and entrepreneurs. Absent financial development, savers can invest directly only in safe, low-return, government-sponsored projects, while financial development can reduce adverse selection, enabling savers to invest in risky (but often more productive) entrepreneurs.

The most effective way of correcting for country and industry characteristics is to use indicator variables, one for each country and industry. Only additional explanatory variables that vary both with industry and country need be included. These are industry j's share in country k of total value added in manufacturing in 1980 and the primary variable of interest, the interaction between industry j's dependence on external financing and financial market development in country k.

The model we want to estimate is then

$$
\begin{aligned}
(1)\quad \text{Growth}_{j,k} &= \text{Constant} + \beta_{1\ldots m}\cdot \text{Country Indicators} \\
&\quad + \beta_{m+1\ldots n}\cdot \text{Industry Indicators} \\
&\quad + \beta_{n+1}\cdot(\text{Industry } j\text{'s share of} \\
&\quad \text{manufacturing in country } k \text{ in } 1980) \\
&\quad + \beta_{n+2}\cdot(\text{External Dependence of} \\
&\quad \text{industry } j\cdot \text{Financial Development} \\
&\quad \text{of country } k) + \varepsilon_{j,k}.
\end{aligned}
$$

Of course, in order to estimate the model, we need appropriate measures of financial development and external dependence. This is what we will examine shortly.

Before proceeding, we point out that our study has one important advantage over recent cross-country empirical studies of growth.[3] That advantage is simply that we make predictions about within-country differences between industries based on an interaction between a country and industry characteristic. Therefore, we can correct for country and industry characteristics in ways that previous studies were unable to correct for, and will be less subject to criticism about an omitted variable bias or model specification.

II. A Measure of Dependence on External Finance

A. *The Proxy for Dependence*

Data on the actual use of external financing is typically not available. But even if it were, it would not be useable because it would reflect the equilibrium between the demand for external funds and its supply. Since the latter is precisely what we are attempting to test for, this information is contaminated. Moreover, we are not aware of systematic studies of the external financing needs of different industries, either cross-sectionally or over time.[4]

We, therefore, have to find some other way of identifying an industry's dependence on external financing. We assume that there is a technological reason why some industries depend more on external finance than others. To the extent that the initial project scale, the gestation period, the cash harvest period, and the requirement for continuing investment differ substantially between industries, this is indeed plausible. Furthermore, we assume that these technological differences persist across countries, so that we can use an industry's dependence on external funds as identified in the United States as a measure of its dependence in other countries. While there are enormous differences in local conditions between countries, all we really need is that statements of the following sort hold: If Pharmaceuticals require a larger initial scale and have a higher gestation period before cash flows are harvested than the Textile industry in the United States, it also requires a larger initial scale and has a higher gestation period in Korea.

B. *How the Proxy Is Calculated*

We start by computing the external financing needs of U.S. companies over the 1980's. We use data from Standard and Poor's *Compustat* (1994) for this. *Compustat* does not contain a representative sample of U.S. firms, because it is limited to publicly traded firms, which are relatively large. Nevertheless, we regard this as an advantage for two reasons. First, in a perfect capital market the supply of funds to firms is perfectly elastic at the proper risk-adjusted rate. In such a market the actual amount of external funds raised by a firm equals its desired amount. In other words, in such an idealized setting, the identification problem does not exist. But capital markets in the United States are among the most advanced in the world, and large publicly traded firms typically face the least frictions in accessing finance. Thus the amount of external finance used by large firms in the United States is likely to be a relatively pure measure of their demand for external finance.[5]

---

[3] See, for example, Roger Kormendi and Philip Meguire (1985), Robert J. Barro (1991), Levine and David Renelt (1992), N. Gregory Mankiw et al. (1992), King and Levine (1993a), and Demirgüç-Kunt and Maksimovic (1996).

[4] Colin Mayer (1990) does look at external financing, but largely at the country level.

A second reason for using a database on listed firms is that disclosure requirements imply that the data on financing are comprehensive. For most of the paper, we will take the amount of external finance used by U.S. firms in an industry as a proxy for the desired amount foreign firms in the same industry would have liked to raise had their financial markets been more developed.

Next, we have to define precisely what we mean by external and internal finance. We are interested in the amount of desired investment that cannot be financed through internal cash flows generated by the same business. Therefore, a firm's dependence on external finance is defined as capital expenditures (*Compustat* # 128) minus cash flow from operations divided by capital expenditures. Cash flow from operations is broadly defined as the sum of cash flow from operations (*Compustat* # 110) plus decreases in inventories, decreases in receivables, and increases in payables.[6] Note that this definition includes changes in the nonfinancial components of net working capital as part of funds from operations. In fact, in certain businesses these represent major sources (or uses) of funds that help a firm avoid (or force it to tap) external sources of funds.[7]

Similarly, the dependence on external equity finance is defined as the ratio of the amount of equity issues (*Compustat* # 108 minus *Compustat* # 115) to capital expenditures. Finally, the investment intensity is the ratio of capital expenditure to net property plant and equipment (*Compustat* # 8).

To make these measures comparable with the industry-level data we have for other countries, we have to choose how to aggregate these ratios over time and across companies. We sum the firm's use of external finance over the 1980's and then divide by the sum of capital expenditure over the 1980's to get the firm's dependence on external finance in the 1980's. This smooths temporal fluctuations and reduces the effects of outliers. To summarize ratios across firms, however, we use the industry median. We do this to prevent large firms from swamping the information from small firms; for instance, we know that IBM's free cash flow does not alleviate possible cash flow shortages of small company firms.

C. *External Dependence for Different Industries*

In Table 1 we tabulate by International Standard Industrial Classification (ISIC) code the fraction of investments U.S. firms financed externally (first column) and the level of capital expenditures divided by net property plant and equipment (second column). We restrict our attention to those manufacturing industries for which we have value-added data from the

---

[5] Even if capital markets are imperfect so that the supply is not perfectly elastic, this methodology provides a reasonable measure of the relative demand for funds provided the elasticity of the supply curve does not change substantially in the cross section. By contrast, in a very imperfect capital market, the relative amount of funds raised may be a function not only of the demand for funds but also of factors that affect supply, such as the availability of collateral.

[6] This item is only defined for cash flow statements with format codes 1, 2, or 3. For format code 7 we construct it as the sum of items # 123, 125, 126, 106, 213, 217.

[7] It could be argued that interfirm trade credit should be viewed as a component of external financing. It is clear how much of trade credit is used to reduce transactions costs and how much is used for financing. trade credit is granted routinely and repaid promptly, usually net trade credit for a firm (accounts receivable payables) is small (see Mitchell A. Petersen and 1997). This may be why trade credit is typically as part of operations in capital budgeting exercises. adhere to this tradition.

United Nations Statistical Division (1993). Drugs and Pharmaceuticals emerge as the industry that uses the most external finance, with Plastics and Computing following close behind. Tobacco, on the other hand, generates the most excess cash flow and has negative external funding needs.

It is common wisdom in the corporate finance literature (though we were hard-pressed to find formal empirical studies of this phenomenon) that there is a life cycle in the pattern of financing for firms; firms are more dependent on external financing early in their life than later. Figure 1 supports the common wisdom. It plots the median financing and investment needs across U.S. firms as a function of the number of years since the initial public offering (IPO). Not surprisingly, in the year of the IPO, firms raise a substantial amount of external funds (especially equity). More interestingly, this continues—albeit on a smaller scale—up to approximately the tenth year. After that period, net equity issues go to zero and the usage of external finance fluctuates around zero. In the third and fourth columns of Table 1, we report the external dependence and capital expenditures for mature companies (firms that were listed for more than ten years), while the fifth and sixth columns are for young companies (firms that were listed for less than ten years).[8] This pattern appears to be fairly standard across different industries, though there are exceptions. All this suggests that very young firms are more dependent on external finance than older firms. This fact will provide an additional test of our hypothesis.

D. *Is the Dependence of U.S. Firms a Good Proxy?*

Much of our analysis rests on dependence of U.S. firms on external finance being a good proxy for the demand for external funds in other countries. We think this is reasonable for four reasons.

First, in a steady-state equilibrium there will not be much need for external funds, as Figure 1 shows. Therefore, much of the demand for external funds is likely to arise as a result of technological shocks that raise an industry's investment opportunities beyond what internal funds can support. To the extent these shocks are worldwide, the need for funds of U.S. firms represents a good proxy.[9]

Second, even if the new investment opportunities generated by these worldwide shocks differ across countries, the amount of cash flow produced by existing firms in a certain industry is likely to be similar across countries. In fact, most of the determinants of ratio of cash flow to capital are likely to be similar worldwide: the level of demand for a certain product, its stage in the life cycle, and its cash harvest period. For this reason, we make sure that our results hold even when we use the amount of internally generated cash, rather than the difference between investments and internally generated funds. We also check that the results hold when we use dependence as measured in Canada, a country which has well-developed capital markets but a very different banking system and industry concentration than the United States. Unfortunately, we do not have access to flow-of-funds data from any other countries, so we cannot venture further afield, but this methodology could, in principle, be used with dependence measured in any country with well-functioning capital markets.

Third, one might argue that the stage of the product life cycle that U.S. firms are in is likely to be different from that of foreign firms. Given that our sample is biased toward developing countries, one might think that the U.S. industry in the 1970's might be a better proxy for the position of developing countries in a product life cycle. For this reason, we also explore the robustness of our results to measuring

---

[8] We required that there be more than one observation in the industry for this variable to be computed. Even with this weak requirement we do not have data for some industries. Most notably there are insufficient young firms in the Tobacco industry.

[9] This amounts to saying that if the invention of personal computers increased the demand for external funds in the U.S. Computer industry, it is likely to increase the need for funds in the Computer industry in other countries as well.

TABLE 1—PATTERN OF EXTERNAL FINANCING AND INVESTMENT ACROSS INDUSTRIES IN THE UNITED STATES DURING THE 1980'S

		All companies		Mature companies		Young companies	
ISIC code	Industrial sectors	External dependence	Capital expenditures	External dependence	Capital expenditures	External dependence	Capital expenditu
314	Tobacco	−0.45	0.23	−0.38	0.24	—	—
361	Pottery	−0.15	0.20	0.16	0.41	−0.41	0.13
323	Leather	−0.14	0.21	−1.33	0.27	−1.53	0.16
3211	Spinning	−0.09	0.16	−0.04	0.19	—	—
324	Footwear	−0.08	0.25	−0.57	0.23	0.65	0.26
372	Nonferrous metal	0.01	0.22	0.07	0.21	0.46	0.24
322	Apparel	0.03	0.31	−0.02	0.27	0.27	0.37
353	Petroleum refineries	0.04	0.22	−0.02	0.22	0.85	0.28
369	Nonmetal products	0.06	0.21	0.15	0.22	−0.03	0.26
313	Beverages	0.08	0.26	−0.15	0.28	0.63	0.26
371	Iron and steel	0.09	0.18	0.09	0.16	0.26	0.19
311	Food products	0.14	0.26	−0.05	0.25	0.66	0.33
3411	Pulp, paper	0.15	0.20	0.13	0.21	0.22	0.20
3513	Synthetic resins	0.16	0.30	−0.23	0.20	0.79	0.45
341	Paper and products	0.18	0.24	0.10	0.23	0.57	0.29
342	Printing and publishing	0.20	0.39	0.14	0.33	0.60	0.41
352	Other chemicals	0.22	0.31	−0.18	0.25	1.35	0.46
355	Rubber products	0.23	0.28	−0.12	0.21	0.50	0.32
332	Furniture	0.24	0.25	0.33	0.17	0.68	0.29
381	Metal products	0.24	0.29	0.04	0.25	0.87	0.34
3511	Basic excluding fertilizers	0.25	0.30	0.08	0.24	0.79	0.29
331	Wood products	0.28	0.26	0.25	0.23	0.34	0.40
384	Transportation equipment	0.31	0.31	0.16	0.28	0.58	0.31
354	Petroleum and coal products	0.33	0.23	0.16	0.26	−0.26	0.22
3843	Motor vehicle	0.39	0.32	0.11	0.33	0.76	0.32
321	Textile	0.40	0.25	0.14	0.24	0.66	0.26
382	Machinery	0.45	0.29	0.22	0.25	0.75	0.33
3841	Ship	0.46	0.43	0.04	0.34	1.05	0.56

TABLE 1—*Continued.*

ISIC code	Industrial sectors	All companies		Mature companies		Young companies	
		External dependence	Capital expenditures	External dependence	Capital expenditures	External dependence	Capital expenditures
390	Other industries	0.47	0.37	−0.05	0.28	0.80	0.49
362	Glass	0.53	0.28	0.03	0.28	1.52	0.33
383	Electric machinery	0.77	0.38	0.23	0.29	1.22	0.46
385	Professional goods	0.96	0.45	0.19	0.33	1.63	0.52
3832	Radio	1.04	0.42	0.39	0.30	1.35	0.48
3825	Office and computing	1.06	0.60	0.26	0.38	1.16	0.64
356	Plastic products	1.14	0.44	—	—	1.14	0.48
3522	Drugs	1.49	0.44	0.03	0.32	2.06	0.47

Notes: This table reports the median level of external financing and capital expenditure for ISIC industries during the 1980's. External dependence is the fraction of capital expenditures not financed with cash flow from operations. Cash flow from operations is broadly defined as the sum of *Compustat* funds from operations (item #110), decreases in inventories, decreases in receivables, and increases in payables. Capital expenditures are the ratio of capital expenditures to net property plan and equipment. Mature companies are firms that have been public for at least ten years; correspondingly, young companies are firms that went public less than ten years ago. The year of going public is the first year in which a company starts to be traded on the NYSE, AMEX, or NASDAQ. All companies is the union of mature and young firms plus firms for which the year of going public could not be determined (firms already traded on NASDAQ in 1972). All the information is obtained from the flow-of-funds data in *Compustat*, except for the SIC code which is obtained from the Center for Research on Securities Prices and then matched with the ISIC code.

the dependence of U.S. firms in the 1970's rather than in the 1980's. We also distinguish between dependence as measured for young firms in the United States (less than ten years from listing) and dependence for old firms (more than ten years from listing).

Last but not least, that we only have a noisy measure of the need for funds creates a bias against finding any interaction between dependence and financial development.

III. Data

A. *Data on Industries*

Data on value added and gross fixed capital formation for each industry in each country are obtained from the *Industrial Statistics Yearbook* database put together by the United Nations Statistical Division (1993). We checked the data for inconsistencies, changes in classification of sectors, and changes in units. The U.N. data is classified by International SIC code. In order to obtain the amount of external finance used by the industry in the United States, we matched ISIC codes with SIC codes.[10] Typically, the three-digit ISIC codes correspond to two-digit SIC codes, while the four-digit ISIC codes correspond to three-digit SIC codes. In order to reduce the dependence on country-specific factors like natural resources, we confine our analysis to manufacturing firms (U.S. SIC 2000–3999).

We would like data on as many countries as possible. The binding constraint is the availability of measures of financial development (specifically the availability of data on

---

[10] Not all the ISIC sectors for which the *Industrial Statistics Yearbook* reports data on value added are mutually exclusive. For example, Drugs (3522) is a subsector of other Chemicals (352). In these cases, the values of the broader sectors are net of the values of the subsectors that are separately reported. We follow this convention both for the data value added and for the financial data from *Compustat.*

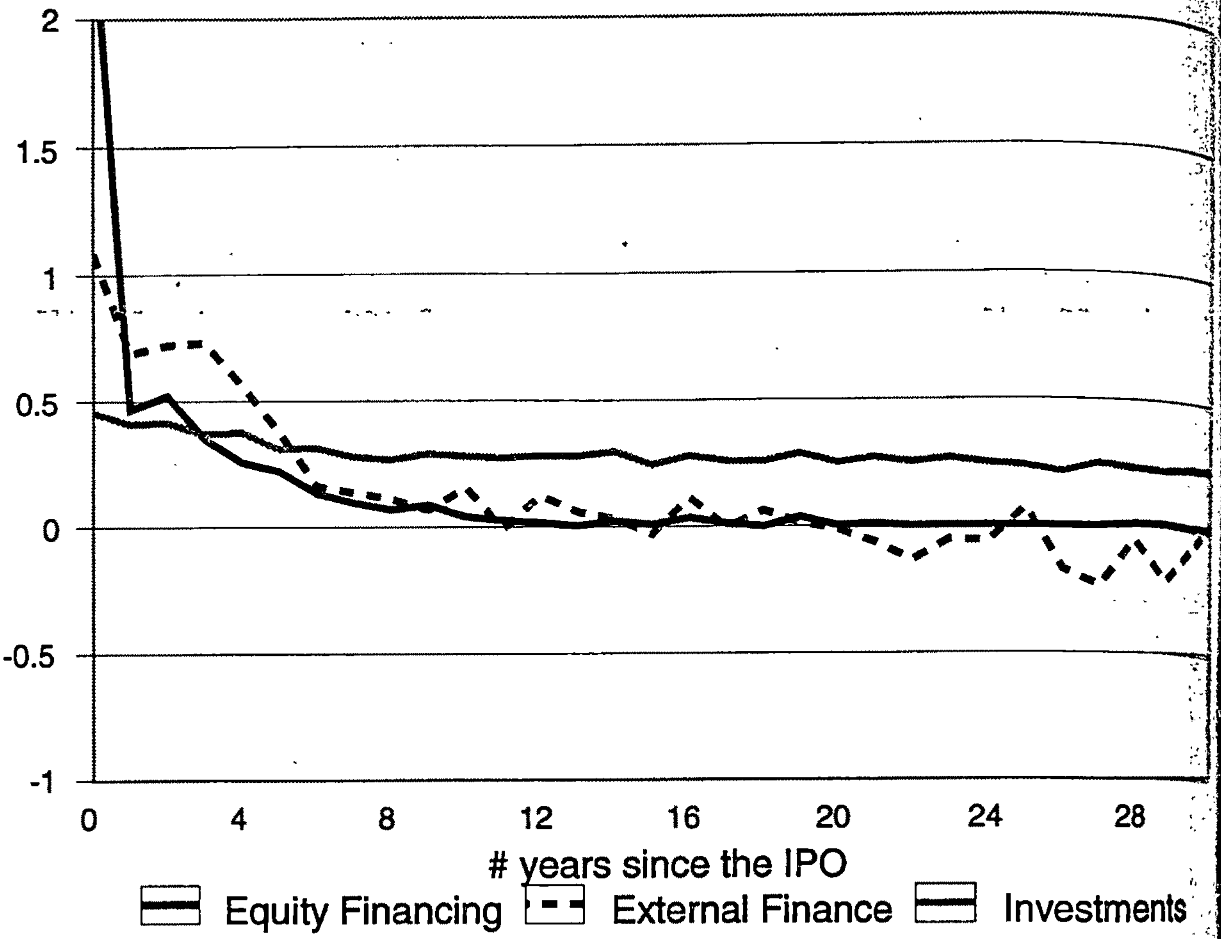


FIGURE 1. LIFE CYCLE OF EXTERNAL FINANCING AND INVESTMENTS

Notes: This graph plots the median level of external financing, equity financing, and investments in the United St across three-digit SIC industries as a function of the number of years since the IPO. External finance is the amount capital expenditures not financed with cash flow from operations, reduction in inventories, or decreases in trade cre Equity finance is the net amount of funds raised through equity issues divided by the amount of investments. Investm is the ratio of capital expenditures to net property, plant, and equipment. The IPO year is defined as the first year in wh a company starts to be traded on the NYSE, AMEX, or NASDAQ. All the information is obtained from the flo of-funds data in *Compustat*, except for the SIC code which is from the Center for Research on Securities Prices.

accounting standards). Since we also wanted data on equity market capitalization, we started with the 55 countries from the International Finance Corporation's (IFC's) *Emerging Stock Markets Factbook.* We dropped countries like Kuwait that did not report a stock market capitalization until the latter half of the 1980's. We could not use Hong Kong and Taiwan because data on these countries are not present in the International Money Fund's (IMF's) *International Financial Statistics* (IFS) volumes. We also dropped countries for which we did not have data from the *Industrial Statistics Yearbook* database that is separated by at least five years (notably, Switzerland). Finally, Thailand is dropped cause the U.N. notes that data from yea year are not comparable. The United State excluded from the analysis because it is benchmark. This leaves us with the 41 co tries in Table 2.

We want to see if financially dependent dustries are likely to be better off in countr with well-developed financial sectors. Th availability of finance affects not just inv ment but also the ability to finance opera and sales through working capital. There the most appropriate measure of an ind being "better off" is the growth in v added for that industry, i.e., the change in of real value added in that industry between 1980 and 1990. Real value added in 1990 is obtained by deflating value added by the Producer Price Index (PPI). For high-inflation countries, spurious differences in value added may be obtained simply because the U.N. data are measured at a different point from the PPI. So, instead, we determine the effective deflator by dividing the growth in nominal value added for the entire manufacturing sector in the U.N. database by the index of industrial production (which measures the real growth rate in industrial production) obtained from the IFS statistics.

B. *Data on Countries*

The Gross Domestic Product, the Producer Price Index, the exchange rate, and the Index of Industrial Production are all obtained from *International Financial Statistics* published by the International Monetary Fund. Whenever a particular series is not available, we use close substitutes—for instance, the Wholesale Price Index if the Producer Price Index is not available. Data on a country's human capital (average years of schooling in population over 25) is obtained from the Barro-Lee files downloaded from the National Bureau of Economic Research web site (see Barro and Jong Wha Lee, 1993).

C. *Measures of Financial Development*

Ideally, financial development should measure the ease with which borrowers and savers can be brought together, and once together, the confidence they have in one another. Thus financial development should be related to the variety of intermediaries and markets available, the efficiency with which they perform the evaluation, monitoring, certification, communication and distribution functions, and the legal and regulatory framework assuring performance. Since there is little agreement on how these are appropriately measured, and even less data available, we will have to make do with crude proxies even though they may miss many of the aspects we think vital to a modern financial system.

The first measure of financial development we use is fairly traditional—the ratio of domestic credit plus stock market capitalization to GDP. We call this the *capitalization ratio.* We obtain stock market capitalization for all countries listed in the *Emerging Stock Markets Factbook* published by the International Finance Corporation, which contains data on developed countries also.[11] Domestic credit is obtained from the IMF's *International Financial Statistics.* Specifically, it is the sum of IFS lines 32a through 32f and excluding 32e. Finally, domestic credit allocated to the private sector is IFS line 32d.

Despite the virtue of tradition, there are concerns with this measure. Unlike domestic credit, stock market capitalization does not reflect the amount of funding actually obtained by issuers. Instead, it reflects a composite of retained earnings, the investing public's perception of the corporate sector's growth prospects, and actual equity issuances. One could argue that the amount of money raised through initial public offerings and secondary offerings is more suitable for our purpose. Unfortunately, these data are not widely available. At the same time, one cannot dismiss the capitalization measure in favor of actual financing too easily. The net amount raised from U.S. equity markets by large firms was negative in the 1980's (see, for example, Rajan and Zingales, 1995). So the actual amount raised may underestimate the importance of the stock market's role in providing price information and liquidity to investors. Market capitalization may be a better measure of the importance of the stock market in this respect. Since we are unsure about whether market capitalization is a reasonable proxy, we will check that the results are robust to redefining the capitalization ratio as the ratio of domestic credit to the private sector to GDP.

The second proxy for financial development we use is the accounting standards in a country. Unlike our first measure, accounting

---

[11] Stock market capitalization is measured at the end of the earliest year in the 1980's for which it is available, while Gross Domestic Product may value flows through the year. This may be a problem in high-inflation countries. We therefore measure GDP as the GDP in constant prices multiplied by the Producer Price Index where the base year for both series is five years before the year of interest.

TABLE 2—FINANCIAL DEVELOPMENT ACROSS COUNTRIES

Country	Accounting standards	Total capitalization over GDP	Domestic credit to private sector over GDP	Per capita income (dollars)
Bangladesh	—	0.20	0.07	121
Kenya	—	0.28	0.20	417
Morocco	—	0.41	0.16	807
Sri Lanka	—	0.44	0.21	252
Pakistan	—	0.53	0.25	290
Costa Rica	—	0.53	0.26	2,155
Zimbabwe	—	1.01	0.30	441
Jordan	—	1.16	0.54	1,109
Egypt	24	0.74	0.21	563
Portugal	36	0.82	0.52	2,301
Peru	38	0.28	0.11	842
Venezuela	40	0.34	0.30	3,975
Colombia	50	0.21	0.14	1,150
Turkey	51	0.35	0.14	1,081
Chile	52	0.74	0.36	2,531
Brazil	54	0.33	0.23	1,650
Austria	54	1.00	0.77	9,554
Greece	55	0.74	0.44	3,814
India	57	0.50	0.24	240
Mexico	60	0.39	0.16	2,651
Belgium	61	0.65	0.29	11,226
Denmark	62	0.56	0.42	12,188
Germany	62	1.08	0.78	12,345
Italy	62	0.98	0.42	6,460
Korea	62	0.63	0.50	1,407
Netherlands	64	0.91	0.60	11,155
Spain	64	1.02	0.76	5,08
Israel	64	1.18	0.67	3,57
Philippines	65	0.46	0.28	72
Japan	65	1.31	0.86	9,91

TABLE 2—*Continued.*

Country	Accounting standards	Total capitalization over GDP	Domestic credit to private sector over GDP	Per capita income (dollars)
France	69	0.70	0.54	11,337
New Zealand	70	0.59	0.19	7,490
South Africa	70	1.51	0.26	2,899
Norway	74	0.63	0.34	13,430
Canada	74	0.98	0.45	10,486
Australia	75	0.82	0.28	9,866
Malaysia	76	1.19	0.48	1,683
Finland	77	0.52	0.48	10,181
U.K.	78	0.78	0.25	9,600
Singapore	78	1.96	0.57	4,661
Sweden	83	0.79	0.42	14,368

Notes: Accounting standards is an index developed by the Center for International Financial Analysis and Research rating the amount of disclosure in annual company reports in each country. Total capitalization to GDP is the ratio of the sum of equity market capitalization (as reported by the IFC) and domestic credit (IFS lines 32a–32f but not 32e) to GDP. Domestic credit to the private sector is IFS line 32d. Per capita income in 1980 is in dollars and is from the IFS.

standards reflect the potential for obtaining finance rather than the actual finance raised. Specifically, the higher the standards of financial disclosure in a country, the easier it will be for firms to raise funds from a wider circle of investors. The Center for International Financial Analysis and Research (CIFAR) creates an index for different countries by rating the annual reports of at least three firms in every country on the inclusion or omission of 90 items. Thus each country obtains a score out of 90 with a higher number indicating more disclosure. The Center for International Financial Analysis and Research, which produces this data, started analyzing balance sheets from 1983 onwards. However, its first comprehensive survey dates from 1990. We will use the accounting standards as measured in this study in much of the paper. The date of the survey raises concerns about endogeneity, but we believe such concerns are small to begin with, and can easily be addressed. First, accounting standards do not change much over time. In 1995, the CIFAR published a study examining how accounting standards had changed since 1983. This study estimated the standards in 1983 and 1990 based on a subset of annual reports, and for a subset of countries that are in the comprehensive 1990 survey. The study finds the mean accounting standards for countries sampled both in 1983 and 1990 is the same at 65. The Wilcoxon signed rank test for equality of distributions fails to reject the equality of the distribution of accounting standards across countries in the two years. Finally, the correlation between the accounting standards in 1983 and 1990 is 0.75.[12] Nevertheless, we will instrument accounting standards with variables that predate the period of growth at which we are looking. Also, we will use the 1983 data to see that the results hold

---

[12] The regression estimates are not sensitive to dropping the few countries such as Denmark and Spain that changed accounting standards substantially.

TABLE 3—SUMMARY STATISTICS

A: Summary Statistics

Variable	Mean	Median	Standard deviation	Minimum	Maximum	Number of observa
Industry's real growth	0.034	0.029	0.099	−0.447	1.000	1242
Industry's growth in number of firms	0.012	0.007	0.071	−0.414	0.759	1073
Industry's growth in average firms' size	0.022	0.026	0.094	−0.536	0.410	1070
Industry's share of total value added	0.016	0.009	0.021	0.000	0.224	1217
Log per capita income in 1980 in dollars	7.870	7.971	1.344	4.793	9.573	41
Average years of schooling	5.900	5.442	2.829	1.681	12.141	41
External finance dependence (all firms)	0.319	0.231	0.319	−0.451	1.492	36
External finance dependence (old firms)	0.010	0.075	0.302	−1.330	0.394	35
External finance dependence (young firms)	0.675	0.673	0.643	−1.535	2.058	34
External finance dependence (1970's)	0.078	0.073	0.188	−0.450	0.542	35
External finance dependence (Canadian firms)	0.427	0.384	0.767	−0.802	3.512	27
Cash flow generated	0.173	0.198	0.112	−0.217	0.331	36
Investment intensity	0.298	0.278	0.095	0.161	0.600	36
Total capitalization over GDP	0.738	0.696	0.375	0.199	1.962	41
Domestic credit to private sector over GDP	0.377	0.302	0.201	0.069	0.856	41
Accounting standards	61.324	62.000	13.238	24.000	83.000	34
Accounting standards (1983)	65.393	68.500	11.426	39.000	81.000	28

B: Correlation Between Measures of External Dependence

	All	Old	Young	1970's	Cash flow	Invest
External finance dependence (all firms)	1.00	—	—	—	—	—
External finance dependence (old firms)	0.46 (0.01)	1.00 —	—	—	—	—
External finance dependence (young firms)	0.72 (0.00)	0.48 (0.00)	1.00 —	—	—	—
External finance dependence (1970's)	0.63 (0.00)	0.42 (0.01)	0.48 (0.00)	1.00 —	—	—
Cash flow generated	−0.91 (0.00)	−0.37 (0.03)	−0.55 (0.00)	−0.50 (0.00)	1.00 —	—
Investment intensity	0.81 (0.00)	0.28 (0.10)	0.64 (0.00)	0.63 (0.00)	−0.60 (0.00)	1.0 —
External finance dependence (Canadian firms)	0.77 (0.00)	0.36 (0.07)	0.58 (0.00)	0.37 (0.07)	−0.78 (0.00)	0.5 (0.0

TABLE 3—*Continued.*

C: Correlation Between Measures of Financial Development

	Total capitalization	Market capitalization	Domestic credit to private sector	Accounting standards	Accounting standards 1983
Market capitalization over GDP	0.79 (0.00)	—	—	—	—
Domestic credit to private sector over GDP	0.67 (0.00)	0.21 (0.18)	1.00 —	—	—
Accounting standards	0.41 (0.02)	0.45 (0.01)	0.25 (0.17)	1.00 —	—
Accounting standards (1983)	0.27 (0.17)	0.39 (0.05)	−0.14 (0.50)	0.68 (0.00)	1.00 —
Per capita income	0.26 (0.09)	0.04 (0.80)	0.48 (0.00)	0.56 (0.00)	0.28 (0.16)

Notes: Industry real growth is the annual compounded growth rate in real value added for the period 1980–1990 for each ISIC industry in each country. The growth in the number of firms is the difference between the log of number of ending-period firms and the log of number of beginning-period firms. The average size of firms in the industry is obtained by dividing the value added in the industry by the number of firms, and the growth in average size is obtained again as a difference in logs. The industry's share of total value added is computed dividing the 1980 value added of the industry by the total value added in manufacturing that year. External dependence is the median fraction of capital expenditures not financed with cash flow from operations for each industry. Cash flow from operations is broadly defined as the sum of *Compustat* funds from operations (items #110), decreases in inventories, decreases in receivables, and increases in payables. External dependence has been constructed using *Compustat* firms between 1980 and 1990, except for Canada where we use *Global Vantage* (Standard & Poor's, 1993) data between 1982 and 1990. Accounting standards is an index developed by the Center for International Financial Analysis and Research ranking the amount of disclosure of companies' annual reports in each country. In Panels B and C the *p*-values are reported in parentheses.

in the subset of countries for which it is available.

Both our measures of financial development, accounting standards and the capitalization ratio, are tabulated for the different countries (see Table 2). While more-developed countries have better accounting standards, there are exceptions. For instance, Malaysia scores as high as Australia or Canada, while Belgium and Germany are in the same league as Korea, the Philippines, or Mexico. Portugal has among the worst accounting standards.

Before we go to the summary statistics, note that for a country's financial development to have any effect on industrial growth in that country we have to assume that firms finance themselves largely in their own country. In other words, only if world capital markets are perfectly integrated can domestic financial development affect a country's growth. There is a wealth of evidence documenting the existence of frictions in international capital markets: the extremely high correlation between a country's savings and its investments (Martin Feldstein and Charles Horioka, 1980), the strong home bias in portfolio investments (Kenneth R. French and James M. Poterba, 1991), and cross-country differences in expected returns (Geert Bekaert and Cambell R. Harvey, 1995). We have little else to say about this assumption other than noting that its failure would weaken the power of our test but not necessarily bias our findings.

Summary statistics and correlations are in Table 3. A number of correlations are noteworthy. First, the financial sector is more developed in richer countries. The correlation of per capita income in 1980 with accounting standards and total capitalization is 0.56 and 0.26 (significant at the 1-percent and 10-percent level, respectively).

Second, the correlation between our capitalization measure of financial development and accounting standards is 0.41 (significant at the 5-percent level for the 33 countries for which we have both data). However, the correlations between accounting standards and the components of capitalization differ. Accounting standards are strongly correlated with equity market capitalization (correlation = 0.45, significant at the 1-percent level) but not with domestic credit (correlation = 0.25, not significant). Domestic credit is credit offered by depository institutions and the central bank. One explanation of the low correlation is perhaps that institutions rely on their own private investigations, and credit from them is little affected by accounting standards. Another possible explanation is that when accounting standards are low, only institutions offer credit. But even though institutions benefit from improvements in accounting standards, other sources of finance become available, and firms substitute away from their traditional sources. We cannot distinguish between these explanations. It will suffice for our purpose that the overall availability of finance, whatever its source, increases with financial development.

IV. Financial Dependence and Growth

A. *Results From the Basic Regression*

1. *Varying Measures of Financial Development.*—Table 4 reports the estimates of our basic specification (1) obtained by using various measures of financial development. Since the specification controls for country-specific effects and industry-specific effects, the only effects that are identified are those relative to variables that vary both cross countries and cross industries. Thus, Table 4 reports only the coefficient of the industry's share of total value added at the beginning of the sample and the coefficient of the interaction between external dependence and different measures of financial development.[13] Since we use U.S. data to identify the external dependence, we drop the United States in all regressions.

We start with total capitalization as the proxy for development. As can be seen in the first column of Table 4, the coefficient estimate for the interaction term is positive and statistically significant at the 1-percent level (throughout the paper, the reported standard errors are robust to heteroskedasticity).[14]

The interaction term is akin to a second derivative. One way to get a sense of its magnitude is as follows. The industry at the 75th percentile of dependence (high dependence) is Machinery. The industry at the 25th percentile (low dependence) is Beverages. The country at the 75th percentile of development as measured by capitalization is Italy, while the country at the 25th percentile is the Philippines. We set the industry's initial share of manufacturing at its overall mean. The coefficient estimate then predicts that Machinery should grow 1.3 percent faster than Beverages annually, and in real terms, in Italy as compared to the Philippines. For comparison, the real annual growth rate is, on average, 3.4 percent per year. So a differential of 1.3 percent is a large number.

For each specification, we compute a similar number which is reported as the *differential in real growth rate* in the last row of each table. Of course, the countries at the 75th and 25th percentile vary with the measure of development, as do the industries at the 75th and 25th percentile with the measure of dependence.

The rest of the columns of the table include different measures of development. We

---

[13] The dependent variable is the average real growth rate over the period 1980–1990. For some countries, however, data availability limits the period. For no country do we have data separated by less than five years. A potential concern is that we measure growth in value added rather than growth in output. Unfortunately, we do not have data for the latter. While we may not capture increases in productivity fully, we see no obvious way in which this should bias our results.

[14] We reduce the impact of outliers by constraining growth between −1 and +1. Three observations are affected. The coefficient estimates for the interaction coefficient are higher and still significant when we do not do this, though the explanatory power of the regression is lower. We also reestimate the same specification after censorizing the 1-percent and 5-percent tails of the growth rate distribution obtaining virtually identical results (except that the explanatory power of the regression is higher).

TABLE 4—INDUSTRY GROWTH AND VARIOUS MEASURES OF DEVELOPMENT

	Financial development measured as					
Variable	Total capitalization	Bank debt	Accounting standards	Accounting standards in 1983	Accounting standards and capitalization	Instrumental variables
Industry's share of total value added in manufacturing in 1980	−0.912 (0.246)	−0.899 (0.245)	−0.643 (0.204)	−0.587 (0.223)	−0.443 (0.135)	−0.648 (0.203)
Interaction (external dependence × total capitalization)	0.069 (0.023)	—	—	—	0.012 (0.014)	—
Interaction (external dependence × domestic credit to private sector)	—	0.118 (0.037)	—	—	—	—
Interaction (external dependence × accounting standards)	—	—	0.155 (0.034)	—	0.133 (0.034)	0.165 (0.044)
Interaction (external dependence × accounting standards 1983)	—	—	—	0.099 (0.036)	—	—
R^2	0.290	0.290	0.346	0.239	0.419	0.346
Number of observations	1217	1217	1067	855	1042	1067
Differential in real growth rate	1.3	1.1	0.9	0.4	1.3	1.0

Notes: The dependent variable is the annual compounded growth rate in real value added for the period 1980–1990 for each ISIC industry in each country. External dependence is the fraction of capital expenditures not financed with internal funds for U.S. firms in the same industry between 1980–1990. The interaction variable is the product of external dependence and financial development. Financial development is total capitalization in the first column, domestic credit to the private sector over GDP in the second column, accounting standards in 1990 in the third column, and accounting standards in 1983 in the fourth column. The sixth column is estimated with instrumental variables. Both the coefficient estimate for the interaction term and the standard error when accounting standards is the measure of development are multiplied by 100. The differential in real growth rate measures (in percentage terms) how much faster an industry at the 75th percentile level of external dependence grows with respect to an industry at the 25th percentile level when it is located in a country at the 75th percentile of financial development rather than in one at the 25th percentile. All regressions include both country and industry fixed effects (coefficient estimates not reported). Heteroskedasticity robust standard errors are reported in parentheses.

private domestic credit to the private sector in the second column, accounting standards in the third column, and accounting standards from the 1983 subsample in the fourth column (for ease of presentation, accounting standards have been divided by 100 in the estimation). The coefficients are uniformly significant at the 1-percent level. The economic magnitudes—as measured by the differential growth rates—are also similar except when development is measured by accounting standards in 1983. The magnitude in column four falls to approximately half of its level otherwise. The explanation for this fall is, perhaps, that the 1983 subsample, being based on just a few companies for each country, introduces significant measurement error.[15]

---

[15] When we instrument this measure (see next paragraph), the coefficient estimate goes up by 50 percent, suggesting the coefficient estimate is biased downwards by measurement error.

In the fifth column, we include both total capitalization and accounting standards. The coefficient for total capitalization is no longer different from zero and its magnitude falls to one-fifth of its level in the first column. Similar results are obtained when we replace total capitalization by domestic credit to the private sector (coefficients not reported). This suggests that accounting standards capture the information about development that is contained in the capitalization measures. For this reason, we will use accounting standards as our measure of development in the rest of the paper. The reader should be assured, however, that the results are qualitatively similar when capitalization measures of development are used.

Because of potential concerns about endogeneity, we will, however, instrument accounting standards with predetermined institutional variables. Rafael La Porta et al. (1996) suggest that the origin of a country's legal system has an effect on the development of a domestic capital market and on the nature of the accounting system. Countries colonized by the British, in particular, tend to have sophisticated accounting standards while countries influenced by the French tend to have poor standards. This suggests using the colonial origin of a country's legal system (indicators for whether it is British, French, German, or Scandinavian) as reported in La Porta et al. as one instrument. Also, countries differ in the extent to which laws are enforced. So we use an index for the efficiency and integrity of the legal system produced by Business International Corporation (a country-risk rating agency) as another instrument. As the sixth column of Table 4 shows, the fundamental interaction becomes even stronger in magnitude when we estimate it using instrumental variables.

Before going further, consider the actual (rather than estimated) effects of development on the growth of specific industries. In Table 5, we summarize for the three least-dependent and three most-dependent industries, the residual growth rate obtained after partialling out industry and country effects. The pattern is remarkable. For countries below the median in accounting standards, the residual growth rate of the three least-dependent industries is positive, while the residual growth rate of the three most-dependent industries is negative. The pattern reverses for countries above the median. Clearly, this suggests no single country or industry drives our results and the rationalized differential in growth rates is systematic and large.

TABLE 5—EFFECT OF FINANCIAL DEVELOPMENT ON ACTUAL GROWTH RATES IN DIFFERENT INDUSTRIES

	Countries below the median in accounting standards	Countries above the median in accounting standards
	Least financially dependent industries	
Tobacco	0.53	−0.60
Pottery	0.25	−0.30
Leather	0.77	−0.77
	Most financially dependent industries	
Drug	−1.11	1.30
Plastics	−0.21	0.21
Computers	−2.00	1.80

Notes: This table reports the mean residual growth rate (in percentage terms) obtained after regressing the annual compounded growth rate in real value added for the period 1980–1990 on industry and country dummies.

2. *Varying Measures of Dependence.*—We now check that our measure of dependence is, indeed, reasonable. We do this in two ways. First, we check that past financing in a country is related to the external dependence of industries in the country. Second, we check that our result is robust to different measures of dependence.

Total capitalization is a (crude) measure of how much finance has been raised in the past in the country. If external dependence is a proxy for an industry's technological need for external finance outside the United States, then countries more specialized in externally dependent industries should have higher capitalization. We calculate the weighted average dependence for each country by multiplying an industry's dependence on external finance by the fraction that the industry contributes to value added in the manufacturing sector in 1980. We then regress total capitalization against weighted average dependence for the 41 countries in the sample. Weighted average dependence is strongly positively correlated with capitalization in 1980 ($\beta = 2.89$, $t = 3.06$). This suggests that our measure of dependence in the United States is related to the external financing used by industry in other countries.[16]

Next, in Table 6 we check that the results are robust to using the external dependence measured for the sample of young firms. Since Figure 1 suggests that most of the demand for external funds is expressed early on in the life of a company, it may be legitimate to expect this to be a better measure of an industry's financial needs. Regardless of how we measure financial development, the interaction effect is positive and statistically significant at the 10-percent level or better, and at the 5-percent level when we use instrumented accounting standards. The magnitude of the coefficient, however, is smaller (roughly a third of the one estimated in Table 4). In part, this reflects the higher level of the external finance raised by young companies. But even when we take this into account (see last row of the table), a difference, albeit smaller, persists. One possible explanation for this result is that young firms are not as important as mature firms in influencing the growth of the industry. We shall return to this issue in Section V, subsection A.

In Table 7, we undertake further robustness checks on our measure of external dependence. While we vary the measure of external dependence, we maintain as a measure of financial development a country's accounting standards, instrumented as above.

In the first column, external dependence is calculated restricting the sample only to mature firms (listed for more than ten years) in the United States. Our interaction variable is positive and statistically significant and the estimated differential growth rate (0.9 percent) is similar to that for the entire sample.

Next, we check whether there is persistence in dependence. If the pattern of financing in the United States in the 1980's is very different from the pattern in the 1970's, it would be unreasonable to expect dependence to carry any information for other countries (especially developing countries that may use older technologies). The raw correlation between an industry's demand for external financing in the 1980's and its demand in the 1970's is 0.63. The coefficient estimate when dependence is measured by the demand for external financing in the 1970's is statistically significant, and the estimated differential growth rate is 0.9 percent.

Finally, it may be that our results derive from the peculiarities of the United States over the 1980's. Our method should work so long as we measure dependence in a country where financial constraints are thought to be small (so that we measure demand not supply). The only other country we have detailed data on flow of funds for is Canada. Canada is very different from the United States along important dimensions. Its banking system is more concentrated as is corporate ownership, and the composition of its industries is different. Nevertheless, the correlation between dependence measured in the United States and dependence measured in Canada is 0.77. As the third column of Table 7 shows, the coefficient estimate when dependence is measured using Canadian data is highly significant. What is especially interesting both in this table and Table 4 is that the economic magnitude of the interaction effect is generally similar despite variation in the measure of dependence and development used.

V. Other Tests

A. *Decomposition of Sources of Growth*

An industry can grow because new establishments are added to the industry or because existing establishments grow in size. The U.N. database also reports the number of establishments in an industry.[17] In our sample, it turns

---

[16] Of course, this raises the possibility of reverse causality, which we will address later.

[17] An establishment is defined as a "unit which engages, under a single ownership or control, in one, or predominantly one, kind of activity at a single location." (*Industrial Statistics Yearbook* p. 4). This definition may not coincide with the legal boundaries of the firm, but is the only one available for such a large cross section of countries.

TABLE 6—INDUSTRY GROWTH AND VARIOUS MEASURES OF DEVELOPMENT USING EXTERNAL DEPENDENCE MEASURED FOR YOUNG FIRMS

	Financial development measured as					
Variable	Total capitalization	Bank debt	Accounting standards	Accounting standards in 1983	Accounting standards and capitalization	Instrume variable
Industry's share of total value added in manufacturing in 1980	−0.911 (0.287)	−0.904 (0.286)	−0.568 (0.234)	−0.616 (0.252)	−0.293 (0.149)	−0.571 (0.233)
Interaction (external dependence × total capitalization)	0.021 (0.012)	—	—	—	−0.004 (0.008)	—
Interaction (external dependence × domestic credit to private sector)	—	0.034 (0.019)	—	—	—	—
Interaction (external dependence × accounting standards)	—	—	0.046 (0.021)	—	0.045 (0.022)	0.05 (0.02
Interaction (external dependence × accounting standards 1983)	—	—	—	0.038 (0.019)	—	—
R^2	0.283	0.283	0.341	0.236	0.415	0.340
Number of observations	1150	1150	1008	808	984	1008
Differential in real growth rate	0.6	0.5	0.4	0.2	0.1	0.5

Notes: The dependent variable is the annual compounded growth rate in real value added for the period 1980–199 each ISIC industry in each country. External dependence is the fraction of capital expenditures not financed with inte funds between 1980–1990 for U.S. firms which went public in the previous ten years belonging to the same ind The interaction variable is the product of external dependence and financial development. Financial development i capitalization in the first column, domestic credit to the private sector over GDP in the second column, acco standards in 1990 in the third column, and accounting standards in 1983 in the fourth column. The sixth colu estimated with instrumental variables. Both the coefficient estimate for the interaction term and the standard error wh accounting standards is the measure of development are multiplied by 100. The differential in real growth rate mea (in percentage terms) how much faster an industry at the 75th percentile level of external dependence grows with re to an industry at the 25th percentile level when it is located in a country at the 75th percentile of financial develop rather than in one at the 25th percentile. All regressions include both country and industry fixed effects (coeffic estimates not reported). Heteroskedasticity robust standard errors are reported in parentheses.

out that *two-thirds* of the growth is spurred by an increase in the average size of establishments, while the remaining third is accounted for by an increase in the number of establishments. The growth in the number of establishments is the log of the number of ending-period establishments less the log of the number of establishments in the beginning of period. The average size of establishments in the industry is obtained by dividing the value added in the industry by the number of establishments, and the growth in average is obtained again as a difference in logs.

Although the definition of establishm provided by the *Industrial Statistics Yearb* does not coincide with the legal definitio a firm, there are three reasons why it is esting to decompose the effect of financial velopment in its effect on the growth i number of establishments and growth in size of the existing establishments. First, s this statistic is often compiled by a diff

TABLE 7—INDUSTRY GROWTH AND VARIOUS MEASURES OF EXTERNAL DEPENDENCE

	External dependence measured using		
Variable	Old firms	Firms in 1970's	Canadian firms
Industry's share of total value added in manufacturing in 1980	−0.625 (0.204)	−0.620 (0.205)	−0.610 (0.235)
Interaction (external dependence × accounting standards)	0.253 (0.063)	0.315 (0.127)	0.065 (0.023)
R^2	0.336	0.334	0.343
Number of observations	1035	1035	802
Differential in real growth rate	0.9	0.9	0.8

Notes: The dependent variable is the annual compounded growth rate in real value added for the period 1980–1990 for each ISIC industry in each country. External dependence is the fraction of capital expenditures not financed with internal funds by firms in the same industry during the 1980's. In the first column this ratio is computed only for companies that have been public for at least ten years. In the second column it is computed for U.S. firms during the 1970's. In the third column it is computed for Canadian firms during the 1980's. Also in the third column, data on U.S. industries are included while data on Canadian industries are dropped. The differential in real growth rate measures (in percentage terms) how much faster an industry at the 75th percentile level of external dependence grows with respect to an industry at the 25th percentile level when it is located in a country at the 75th percentile of financial development rather than in one at the 25th percentile. All regressions are estimated using instrumental variables and include both country and industry fixed effects (coefficient estimates not reported). Heteroskedasticity robust standard errors are reported in parenthesis.

body in a country from the one that produces the value-added data, this test provides an independent check on our results.[18] Second, the creation of new establishments is more likely to require external funds, while the expansion of existing establishments can also use internal funds. Thus, the effect of financial development should be more pronounced for the first than for the second. Finally, the growth in the number of establishments is more likely to be generated by new firms than the growth in the size of the existing establishments. Thus, the growth in the number of establishments should be more sensitive to the external dependence measured using young firms in the United States.

We then estimate the basic regression with growth in number of establishments and growth in average size as dependent variables. As Table 8 indicates, the interaction variable is statistically significant only when explaining the growth in the number of establishments. More important, the differential in growth rate suggested by the estimate is twice as large in the second column (the regression with growth in numbers as the dependent variable) as in the first column (the regression with growth in average size as the dependent variable).

This finding that the development of financial markets has a disproportional impact on the growth of new establishments is suggestive. Financial development could indirectly influence growth by allowing new ideas to develop and challenge existing ones, much as Schumpeter argued.

Recall that in the previous section, we found that the dependence of young firms was of lower importance (both statistical and economic) than the dependence of mature firms in explaining the relative growth of industries. One explanation is that the dependence of young firms in the United States is an accurate measure of the needs of new firms in that industry elsewhere, but only a noisy measure of the dependence of all firms. This seems to be the case. When dependence is measured for

---

[18] The disadvantage is that the industry classification used by the body compiling the number of firms may differ from the industry classification used by the body compiling the value-added data, resulting in an increase in noise.

TABLE 8—GROWTH IN AVERAGE SIZE AND NUMBER OF ESTABLISHMENTS

	External dependence measured using				
	All firms		Young firms		Mature firms
Variable	Growth average size	Growth number	Growth average size	Growth number	Growth average size
Industry's share of total value added in manufacturing in 1980	−0.620 (0.217)	−0.312 (0.154)	−0.635 (0.256)	−0.252 (0.179)	−0.624 (0.220)
Interaction (external dependence × accounting standards)	0.051 (0.043)	0.115 (0.037)	−0.021 (0.029)	0.078 (0.024)	0.125 (0.055)
R^2	0.498	0.314	0.500	0.302	0.492
Number of observations	951	975	899	922	923
Differential in real growth rate	0.3	0.7	−0.2	0.6	0.4

Notes: The average size of establishments in the industry is obtained by dividing the value added in the industry by number of establishments, and the growth in average size is obtained as a difference in logs between average size in and average size in 1980. The growth in the number of establishments is the log of the number of establishments in less the log of the number of establishments in 1980. The differential in real growth rate measures (in percentage how much faster an industry at the 75th percentile level of external dependence grows with respect to an industry 25th percentile level when it is located in a country at the 75th percentile of financial development rather than in the 25th percentile. All regressions are estimated using instrumental variables and include both country and industry effects (coefficient estimates not reported). Heteroskedasticity robust standard errors are reported in parentheses.

young firms, the interaction coefficient has a positive, statistically significant effect on the growth in the number of establishments, but a negative (and statistically insignificant) effect on the growth of the average size of existing establishments (third and fourth columns); when dependence is measured for mature firms, the interaction coefficient has a positive, statistically significant effect on both.

Since most of growth in value added is generated by an increase in the average size of existing establishments, the most appropriate measure of external dependence seems to be one that includes both the needs of new firms as well as the needs of existing firms. This is why in the rest of the paper we will use external dependence measured across all firms.

B. *Is the Interaction a Proxy for Other Variables?*

Do external dependence or financial development proxy for something else? In principle, there is a long list of sources of comparative advantage that may dictate the presence, absence, or growth of industries in a country. O results, though, cannot be explained unless dependence of industries on this source comparative advantage is strongly correl with their dependence on external funding financial development is a good proxy for source of comparative advantage. We rule two such possibilities below.

Industries that are highly dependent on ternal finance—for example, Drugs and Ph maceuticals—could also be dependent human capital inputs. To the extent that fi cial market development and the availab of human capital are correlated, the obse interaction between external dependence financial development may proxy for the teraction between human capital depend and the availability of trained human ca To check this, we include in the basic re sion an interaction between the industry pendence on external finance and a meas the country's stock of human capital (av years of schooling in population over th of 25). If the conjecture is true, the coeffi of the financial development interaction

TABLE 9—ROBUSTNESS CHECKS

Variable	Human capital	Economic development	Above median	Below median
Industry's share of total value added in manufacturing in 1980	−0.386 (0.137)	−0.422 (0.134)	−0.437 (0.178)	−6.079 (1.932)
Interaction (external dependence × accounting standards)	0.191 (0.072)	0.149 (0.055)	0.161 (0.065)	0.161 (0.066)
Interaction 2 (external dependence × average years of schooling)	−0.002 (0.003)	—	—	—
Interaction 3 (external dependence × log of per capita income in 1980)	—	0.000 (0.005)	—	—
R^2	0.413	0.418	0.548	0.390
Number of observations	1006	1042	522	545
Differential in real growth rate	1.0	0.9	0.9	1.0

Note: The dependent variable is the annual compounded growth rate in real value added for the period 1980–1990 for each ISIC industry in each country. The first column adds to the basic specification the interaction between external dependence and a country's human capital. The second column adds to the basic specification the interaction between external dependence and a country's level of economic development (log per capita income). The third column estimates the basic specification for industries that in 1980 were above the median industry in terms of the fraction they accounted for of value added in the manufacturing sector. The fourth column estimates the basic specification for industries that in 1980 were below the median industry in terms of the fraction they accounted for of value added in the manufacturing sector. The differential in real growth rate measures (in percentage terms) how much faster an industry at the 75th percentile level of external dependence grows with respect to an industry at the 25th percentile level when it is located in a country at the 75th percentile of financial development rather than in one at the 25th percentile. All regressions are estimated using instrumental variables and include both country and industry fixed effects (coefficient estimates not reported). Heteroskedasticity robust standard errors are reported in parentheses.

should fall substantially. As the coefficient estimates in the first column of Table 9 show, the coefficient on the human capital interaction term is small and not statistically significant, while the financial development interaction increases somewhat. This suggests that financial dependence is not a proxy for the industry's dependence on human capital.

Another possibility is that lower dependence on external financing in the United States simply reflects the greater maturity of the industry. An influential view of the development process is that as technologies mature, industries using those technologies migrate from developed economies to developing economies (see, for example, Rudiger Dornbusch et al., 1977). Since developing countries are more likely to have underdeveloped financial markets, the interaction effect we document may simply reflect the stronger growth of mature technologies in underdeveloped countries.

We already have results suggesting this cannot be the entire explanation. The interaction effect is present even when dependence is measured only for young firms in the United States. Furthermore, we can test if financial development is really a proxy for economic development in the regression. We include in the basic regression the interaction between the industry's dependence on external finance and the log per capita GDP for the country, in addition to our usual interaction term. As seen in the second column of Table 9, the coefficient of the interaction term falls from 0.165 (in the basic regression) to 0.149 but is still statistically and economically significant. The interaction between financial dependence and log per capita income is close to zero and not significant. The results do not suggest financial dependence is a proxy for technological maturity.

C. *Other Explanations: Reverse Causality*

Thus far, we have taken the state of financial markets as predetermined and exogenous. An alternative explanation of the development of financial markets is that they arise to accommodate the financing needs of finance-hungry industries.

The argument is as follows. Suppose there are some underlying country-specific factors or endowments (such as natural resources) that favor certain industries (such as Mining) that happen to be finance hungry. Then, countries abundant in these factors should experience higher growth rates in financially dependent industries and—as a result—should develop a strong financial market. If these factors persist, then growth rates in financially dependent sectors will persist and we will observe the significant interaction effect. But here it will result from omitted factors than any beneficial effect of finance.

On the one hand, the lack of persistence in country growth over periods of decades (see William Easterly et al., 1993) and the low correlation of sectoral growth across decades (see Peter Klenow, 1995) suggest that this should not be a major concern. On the other hand, our finding that capitalization is higher when the weighted average dependence of industries in the country is high indicates the argument is not implausible.

The results we already have should reduce concerns about reverse causality. By restricting the sample to manufacturing firms, we have reduced the influence of availability of natural resources. More important, the measure of financial development we use—accounting standards—is instrumented with predetermined variables that are unlikely to be correlated with omitted factors driving the growth of industries dependent on external finance. In fact, it should be less correlated with past financing than the capitalization measure, yet it explains future relative growth rates better.

However, we can also test the argument more directly. If an industry has a substantial presence in a particular country, it is logical that the country has the necessary resources and talents for the industry. So by further restricting the sample to industries that are above the median size in the country in 1980, we reduce the problem of differences in growth stemming from differences in endowment. When we estimate the regression with this smaller sample (third column of Table 9), the interaction coefficient is virtually unchanged.

One way to make sense of all our findings without reverse causality driving the results is that financial markets and institutions may develop to meet the needs of one set of industries, but then facilitate the growth of another younger group of industries. Alfred D. Chandler, Jr. (1977) suggests this is, in fact, what happened in the United States. The financial sector, especially investment banks and the corporate bond market, developed to meet the financing needs of railroads in the mid-nineteenth century. The financial infrastructure was, therefore, ready to meet the financing needs of industrial firms as they started growing in the latter half of the nineteenth century. Similarly, Goldsmith (1985, 2) based on a study of the balance sheets of 20 countries writes: "The creation of a modern financial superstructure, not in its details but in its essentials, was generally accomplished at a fairly early stage of a country's economic development."

Again, we can test this possibility more directly. We estimate the effect of financial development only for industries that are small to start out with, and are unlikely to be responsible for the state of development of the financial markets. So we estimate the basic regression for industries that in 1980 were lower than the median size in their respective countries. The coefficient of the interaction term is again unchanged (see column four of Table 9) even for these industries for which the economy's financial development is largely predetermined. We conclude that reverse causality is unlikely to explain our results.

D. *Other Explanations: Investment and Cost of Capital*

Investment opportunities in different industries may be very different. For instance,

TABLE 10—CASH FLOW AND INVESTMENTS

Variable	Cash flow intensiveness	Investment intensiveness	Both	Both measured for 1980
Industry's share of total value added in manufacturing in 1980	−0.588 (0.201)	−0.653 (0.205)	−0.639 (0.205)	−0.639 (0.207)
Interaction (internal cash flow × financial development)	0.482 (0.153)	—	−0.261 (0.196)	−0.595 (0.295)
Interaction 2 (investment intensiveness × accounting standards)	—	0.623 (0.221)	0.443 (0.283)	0.800 (0.299)
R^2	0.343	0.345	0.345	0.344
Number of observations	1067	1067	1067	1035
Differential in real growth rate	−0.7	1.4	0.5	1.6

Notes: The dependent variable is the annual compounded growth rate in real value added for the period 1980–1990 for each ISIC industry in each country. Internal cash flow is the ratio of cash flow from operations broadly defined (see text) to net property plant and equipment for U.S. firms in the same industry. Investment intensity is the ratio of capital expenditures to property plant and equipment for U.S. firms in the same industry. The fourth column uses the cash flow intensity and the investment intensity measured for the year 1980. The differential in real growth rate measures (in percentage terms) how much faster an industry at the 75th percentile level of external dependence grows with respect to an industry at the 25th percentile level when it is located in a country at the 75th percentile of financial development rather than in one at the 25th percentile. All regressions are estimated using instrumental variables and include both country and industry fixed effects (coefficient estimates not reported). Heteroskedasticity robust standard errors are reported in parentheses.

Tobacco industry in the United States uses negative external finance (see Table 1) partly because investment opportunities in the Tobacco industry are small relative to the cash flows the industry generates. It may be that our measure of dependence on external finance proxies primarily for the investment intensity of a particular industry. Furthermore, the development of the financial sector may proxy for the overall cost of capital in that country (rather than the cost of external funds). The interaction effect then indicates that capital intensive firms grow faster in an environment with a lower cost of capital. Though this is a legitimate channel through which the financial sector influences growth, we are also interested in a different channel where the reduction in the incremental cost of external funds facilitates growth.

If investment intensity were all that mattered, and external finance and internal finance were equally costly, the cash internally generated by industries would be irrelevant in countries that are more financially developed. All that mattered would be the size of the required investment and the cost of capital. By contrast, if there is a wedge between the cost of internal and external finance which narrows as the financial sector develops, industries generating lots of internal cash should grow relatively faster in countries with a poorly developed financial sector. As indicated in the first column of Table 10, they do. This is consistent with financial development reducing the cost of external finance. Of course, as is to be expected with both the "cost of capital" and "cost of external capital" hypotheses, industries that invest a lot also grow faster in countries with more-developed financial markets (second column). Unfortunately, when both interactions are introduced in the same regression, the coefficients are measured very imprecisely because of multicollinearity (cash flow intensity and investment intensity have a correlation of 0.73); so neither is statistically different from zero. However, the coefficient on cash flows is still negative and sizeable (accounting for a real growth rate differential of about 0.4 percent per year).

Multicollinearity results from our aggregating cash flows and investments over a decade.[19] Therefore, we estimate the same regression using a measure of cash flow intensity and investment intensity measured for just one year (rather than an entire decade). In the fourth column we report the estimates obtained by using the 1980 measures of cash flow and investment. Both the cash flow intensity and the investment intensity are statistically significant at the 5-percent level. We estimated (but do not report) the same regression using a 1985 measure and a 1990 measure. In both cases the results are similar and both coefficients are statistically significant at the 5-percent level.

VI. Conclusion

We develop a new methodology in this paper to investigate whether financial-sector development has an influence on industrial growth. In doing so, we partially circumvent some of the problems with the existing cross-country methodology highlighted by Mankiw (1995). First, it is difficult to interpret observed correlations in cross-country regressions in a causal sense. Here, we push the causality debate one step further by finding evidence for a channel through which finance theoretically influences growth. Also, since we have multiple observations per country, we can examine situations where the direction of causality is least likely to be reversed. A second problem with the traditional methodology is that explanatory variables are multicollinear and are measured with error. The combination of these two problems may cause a variable to appear significant when it is merely a proxy for some other variable measured with error. As a result, observed correlations can be misleading. By looking at interaction effects (with country and industry indicators) rather than direct effects, we reduce the number of variables that we rely on, as well as the range of possible alternative explanations. Third, there is the problem of limited degrees of freedom—there are fewer than 200 countries on which myriad theories have to be tested. Our proach partially alleviates this problem by ploiting within-country variation in the d Our methodology, may have wider app tions, such as testing the existence of chan through which human capital can aff growth.

Apart from its methodological contributi this paper's findings may bear on three diffe ent areas of current research. First, th suggest that financial development has a s stantial supportive influence on the rate of ec nomic growth and this works, at least part by reducing the cost of external finance to f nancially dependent firms. We should add th there is no contradiction when the lack of pe sistence of economic growth (Easterly et a 1993) is set against the persistence of financ development. Other factors may cause (po tially serially uncorrelated) changes in a co try's investment opportunity set. Finance m simply enable the pursuit of these opport ties, and thereby enhance long-run grow The paper does, however, suggest that fin cial development may play a particularly be eficial role in the rise of new firms. If th firms are disproportionately the source ideas, financial development can enhance novation, and thus enhance growth in indir ways.

Second, in the context of the literature financial constraints, this paper provides fir evidence that financial market imperfecti have an impact on investment and growth.

Finally, in the context of the trade litera the findings suggest a potential explanatio the pattern of industry specialization acr countries. To the extent that financial-ma development (or the lack thereof) is de mined by historical accident or governm regulation, the existence of a well-develo market in a certain country represents a so of comparative advantage for that country industries that are more dependent on ext finance. Similarly, the costs imposed by a of financial development will favor incum firms over new entrants. Therefore, the l of financial development can also be a f in determining the size composition of a dustry as well as its concentration. These sues are important areas for future resea

---

[19] Early investments will generate later cash flows resulting in the correlation. Aggregating over a decade, however, will still give a reasonable estimate of the average demand for external funds even though it tells us less about the components.

REFERENCES

Barro, Robert J. "Economic Growth in a Cross Section of Countries." *Quarterly Journal of Economics*, May 1991, *106*(2), pp. 407–43.

Barro, Robert J. and Lee, Jong Wha. "International Comparisons of Educational Attainment." *Journal of Monetary Economics*, December 1993, *32*(3), pp. 363–94.

Bekaert, Geert and Harvey, Campbell R. "Time-Varying World Market Integration." *Journal of Finance*, June 1995, *50*(2), pp. 403–44.

Bencivenga, Valerie R. and Smith, Bruce D. "Financial Intermediation and Endogenous Growth." *Review of Economic Studies*, April 1991, *58*(2), pp. 195–209.

Boyd, John H. and Smith, Bruce D. "The Co-Evolution of the Real and Financial Sectors in the Growth Process." *World Bank Economic Review*, May 1996, *10*(2), pp. 371–96.

Cameron, Rondo. *Banking in the early stages of industrialization*. New York: Oxford University Press, 1967.

Chandler, Alfred D., Jr. *The visible hand: The managerial revolution in American business*. Cambridge, MA: Harvard University Press, 1977.

Demirgüç-Kunt, Asli and Maksimovic, Vojislav. "Financial Constraints, Uses of Funds and Firm Growth: An International Comparison." Working paper, World Bank, 1996.

Diamond, Douglas W. "Reputation Acquisition in Debt Markets." *Journal of Political Economy*, August 1989, *97*(4), pp. 828–62.

Dornbusch, Rudiger; Fischer, Stanley and Samuelson, Paul A. "Comparative Advantage, Trade and Payments in A Ricardian Model with a Continuum of Goods." *American Economic Review*, December 1977, *67*(5), pp. 823–39.

Easterly, William; Kremer, Michael; Pritchett, Lant and Summers, Lawrence H. "Good Policy or Good Luck? Country Growth Performance and Temporary Shocks." *Journal of Monetary Economics*, December 1993, *32*(3), pp. 459–83.

Feldstein, Martin and Horioka, Charles. "Domestic Saving and International Capital Flows." *Economic Journal*, June 1980, *90*(358), pp. 314–29.

French, Kenneth R. and Poterba, James M. "Investor Diversification and International Equity Markets." *American Economic Review*, May 1991 (*Papers and Proceedings*), *81*(2), pp. 222–26.

Goldsmith, Raymond, W. *Financial structure and development*, New Haven, CT: Yale University Press, 1969.

______. *Comparative national balance sheets*. Chicago: University of Chicago Press, 1985.

Greenwood, Jeremy and Jovanovic, Boyan. "Financial Development, Growth, and the Distribution of Income." *Journal of Political Economy*, October 1990, *98*(5), pp. 1076–107.

International Finance Corporation. *Emerging stock markets factbook*. Washington, DC: International Finance Corporation, various years.

International Monetary Fund. *International financial statistics*. Washington, DC: International Monetary Fund, various years.

Jayaratne, Jith and Strahan, Philip E. "The Finance-Growth Nexus: Evidence from Bank Branch Deregulation." *Quarterly Journal of Economics*, August 1996, *111*(3), pp. 639–70.

Jensen, Michael C. and Meckling, William R. "Theory of the Firm: Managerial Behavior, Agency Costs and Capital Structure." *Journal of Financial Economics*, October 1976, *3*(4), pp. 305–60.

King, Robert G. and Levine, Ross. "Finance and Growth: Schumpeter Might Be Right." *Quarterly Journal of Economics*, August 1993a, *108*(3), pp. 713–37.

______. "Finance, Entrepreneurship, and Growth." *Journal of Monetary Economics*, December 1993b, *32*(3), pp. 513–42.

Klenow, Peter. "Ideas vs. Rival Human Capital: Industry Evidence on Growth Models." Mimeo, University of Chicago, 1995.

Kormendi, Roger and Meguire, Philip. "Macroeconomic Determinants of Growth: Cross-Country Evidence." *Journal of Monetary Economics*, September 1985, *16*(2), pp. 141–63.

La Porta, Rafael; Lopez de Silanes, Florencio; Shleifer, Andrei and Vishny, Robert. "Law and Finance." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5661, 1996.

Levine, Ross. "Financial Development and Economic Growth: Views and Agenda." *Journal of Economic Literature*, June 1997, *35*(2), pp. 688–726.

Levine, Ross and Renelt, David. "A Sensitivity Analysis of Cross-Countries Growth Regressions." *American Economic Review*, September 1992, *82*(4), pp. 942–63.

Levine, Ross and Zervos, Sara. "Stock Markets and Economic Growth." *American Economic Review*, June 1998, *88*(3), pp. 537–58.

Mankiw, N. Gregory. "The Growth of Nations." *Brookings Papers on Economic Activity*, 1995, (1), pp. 275–310.

Mankiw, N. Gregory; Romer, David and Weil, David. "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, May 1992, *152*(2), pp. 407–37.

Mayer, Colin. "Financial Systems, Corporate Finance, and Economic Development," in R. Glenn Hubbard, ed., *Asymmetric information, corporate finance and investment*. Chicago: University of Chicago Press, 1990, pp. 307–32.

Myers, Stewart C. and Majluf, Nicholas S. "Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have." *Journal of Financial Economics*, June 1984, *13*(2), pp. 187–221.

Obstfeld, Maurice. "Risk Taking, Global Diversification, and Growth." *American Economic Review*, December 1994, *84*(5), 1310–29.

Petersen, Mitchell A. and Rajan, Raghuram "Trade Credit: Theories and Evidence." *Review of Financial Studies*, Fall 199 *10*(3), pp. 661–91.

Rajan, Raghuram R. and Zingales, Luigi. "W Do We Know About Capital Structur Some Evidence From International Data." *Journal of Finance*, December 199 *50*(5), pp. 1421–60.

Robinson, Joan. "The Generalization of t Generat Theory." *The rate of interest a other essays*. London: Macmillan, 1952, p 67–142.

Saint-Paul, Gilles. "Technological Choice, Financial Markets and Economic Development." *European Economic Review*, Ma 1992, *36*(4), pp. 763–81.

Schumpeter, Joseph A. *A theory of econo development*. Cambridge, MA: Harvar University Press, 1911.

Standard & Poor's. *Global vantage*. New Yo McGraw-Hill, 1993.

______. *Compustat services*. New Yo McGraw-Hill, 1994.

Stiglitz, Joseph E. and Weiss, Andrew. "Cre Rationing in Markets with Imperfect Information." *American Economic Review*, Ju 1981, *71*(3), pp. 393–410.

United Nations, Statistical Division. *Industri statistics yearbook*, Vol. 1. New Yo United Nations, 1993.

Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks

By STEPHEN MORRIS AND HYUN SONG SHIN*

Even though self-fulfilling currency attacks lead to multiple equilibria when fundamentals are common knowledge, we demonstrate the uniqueness of equilibrium when speculators face a small amount of noise in their signals about the fundamentals. This unique equilibrium depends not only on the fundamentals, but also on financial variables, such as the quantity of hot money in circulation and the costs of speculative trading. In contrast to multiple equilibrium models, our model allows analysis of policy proposals directed at curtailing currency attacks. (*JEL* F31, D82)

Speculative attacks are sometimes triggered without warning, and without any apparent change in the economic fundamentals. Commentators who have attempted to explain episodes of speculative crises have pointed to the self-fulfilling nature of the belief in an imminent speculative attack. If speculators believe that a currency will come under attack, their actions in anticipation of this precipitate the crisis itself, while if they believe that a currency is not in danger of imminent attack, their inaction spares the currency from attack, thereby vindicating their initial beliefs.[1]

However, merely pointing to the self-fulfilling nature of beliefs leaves open a number of crucial questions. First, such an account leaves unexplained the actual *onset* of an attack when it occurs. By most accounts, both the European Exchange Rate Mechanism (ERM) and the Mexican peso were "ripe" for attack for some time before the crises that brought them down in the early 1990's—at least two years in Europe and perhaps a year in Mexico (Barry Eichengreen and Charles Wyplosz, 1993; Rudiger Dornbusch and Alejandro M. Werner, 1994). At any point in those periods, concerted selling by speculators would have raised the costs of maintaining the exchange rate sufficiently high that the monetary authorities would have been forced to abandon the parity. Why did the attacks happen when they did? Conventional accounts resort to forces that operate outside the theoretical model in trying to explain the shift of expectations that precipitated the attack. Some of these informal accounts are more persuasive than others, but none can be fully compelling as a theoretical model of the onset of a currency crisis.

Secondly, merely pointing to the self-fulfilling nature of beliefs leaves open the policy issues associated with curbing speculative attacks. For example, it is often argued that increased capital mobility induced by lower transaction costs increases the likelihood of currency crises, and that judicious "throwing of sand" into the excessively well-oiled wheels of international finance will play a role in curbing speculative attacks (Eichengreen et al., 1995). However, an account that merely points to the self-fulfilling nature of beliefs cannot contribute to this debate, since the question of how the beliefs are determined is beyond the scope of such an account. Any argument for or against the proposal has to resort to forces outside the formal theory.

---

* Morris: Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104; Shin: Nuffield College, Oxford OX1 1NF, England. We are grateful to Marcus Miller for encouraging us to pursue this topic and to Stephen Coate and John Duffill for advice during the preparation of this paper. We thank Anne Sibert for her comments as discussant at the CEPR conference on "Origins and Management of Financial Crises." The first author acknowledges the support of the Alfred P. Sloan Foundation. The second author acknowledges the support of the U.K. Economic and Social Research Council under Grant No. R000221556. Finally, we thank two anonymous referees for comments on an earlier version.

[1] This theme is explored in Robert P. Flood and Peter M. Garber (1984) and Maurice Obstfeld (1986, 1994, 1996). A closely related literature explores self-fulfilling [illegible] crises; see Guillermo A. Calvo (1988), Jonathan [illegible] and Raquel Fernandez (1995), and Harold L. Cole and Timothy J. Kehoe (1996).

We argue here that the apparent multiplicity of equilibria associated with self-fulfilling beliefs is the consequence of assuming too simple a picture of the role played by information in a speculative episode. Although market participants each have a window on to the world, the imperfect nature of such a vantage point generates a failure of common knowledge of the fundamentals. Thus, everyone may know that the fundamentals are sound, but it may not be that everyone knows that everyone knows this. Still higher orders of such uncertainty may be relevant. Uncertainty about other participants' beliefs is crucial to a speculative episode, since the onset of a speculative attack relies to a large extent on coordinated behavior of speculators. We propose a more realistic modelling of the information structure underlying speculative situations that enables us to pinpoint the forces which determine the onset of a speculative attack. When speculators observe an independent noisy signal of the state of fundamentals, common knowledge of the fundamentals no longer holds. This is enough to induce a unique equilibrium of the model, in which there is a critical state below which attack always occurs and above which attack never occurs.[2] The value of this critical state depends on financial variables such as the mass of speculators and the transaction costs associated with attacking the currency. As a consequence, we can say something about how this unique outcome depends on the parameters of the problem, such as the costs of speculation, the underlying strength of the economy, and the size of the pool of hot money in circulation.

Information plays a subtle role in speculative crises. What is important is not the amount of information, per se, but rather how public and transparent this information is. If market participants are well informed about the funda[illegible]tals, but they are unsure of the inform[illegible] received by other participants, and hence [illegible] sure of the beliefs held by others, specul[illegible] attacks may be triggered even though every[illegible] knows that the fundamentals are sound. [illegible] analysis highlights the importance of the tr[illegible] parency of the conduct of monetary policy [illegible] its dissemination to the public.

I. Model

Our model is concerned with the strat[illegible] interaction between the government and [illegible] group of speculators in the foreign exch[illegible] market that takes place against the backd[illegible] of a competitive market for foreign exchan[illegible] The economy is characterized by a state [illegible] fundamentals θ, which we assume to be [illegible] formly distributed over the unit interval $[0,1]$. The exchange rate in the absence of gov[illegible] ment intervention is a function of θ, and [illegible] given by $f(\theta)$. We assume that f is stri[illegible] increasing in θ, so that higher values of θ co[illegible] respond to "stronger fundamentals."

The exchange rate is initially pegged by [illegible] government at e^*, where $e^* \geq f(\theta)$ for all [illegible] Facing the government is a continuum of spe[illegible] ulators who may take one of two actions. [illegible] speculator may either attack the currency b[illegible] selling short one unit of the currency, or [illegible] frain from doing so. There is a cost $t > 0$ [illegible] sociated with short-selling. If a specul[illegible] short-sells the currency and the governm[illegible] abandons the exchange rate peg, then the p[illegible] off at state θ by attacking the currency is gi[illegible] by the fall in the exchange rate minus [illegible] transaction cost. Thus the speculator's pay[illegible] is $e^* - f(\theta) - t$. If the government def[illegible] the peg, then the speculator pays the trans[illegible] tion cost, but has no capital gain. Thus his p[illegible] off is $-t$. If the speculator chooses not [illegible] attack the currency, his payoff is zero.

The government derives a value $v > 0$ fr[illegible] defending the exchange rate at the peg[illegible] level, but also faces costs of doing so. The c[illegible] of defending the peg depends on the sta[illegible] the fundamentals, as well as the proportio[illegible] speculators who attack the currency. We [illegible] note by $c(\alpha, \theta)$ the cost of defending the [illegible] if proportion α of the speculators attack [illegible] currency at the state θ. The government's p[illegible]

---

[2] Andrew Postlewaite and Xavier Vives (1987) discuss the role of information in selecting among multiple equilibria in the closely related context of bank runs. Our analysis builds on and extends the earlier incomplete information game analysis of Hans Carlsson and Eric E. van Damme (1993) and Morris et al. (1995).

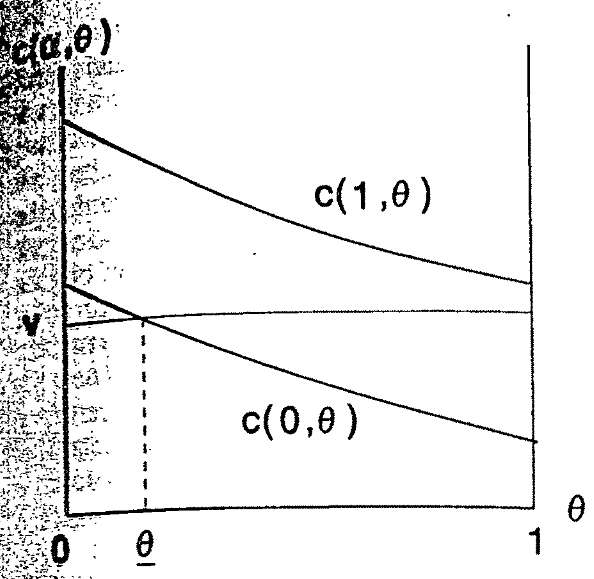


FIGURE 1. COST AND BENEFIT TO THE GOVERNMENT IN MAINTAINING THE CURRENCY PEG

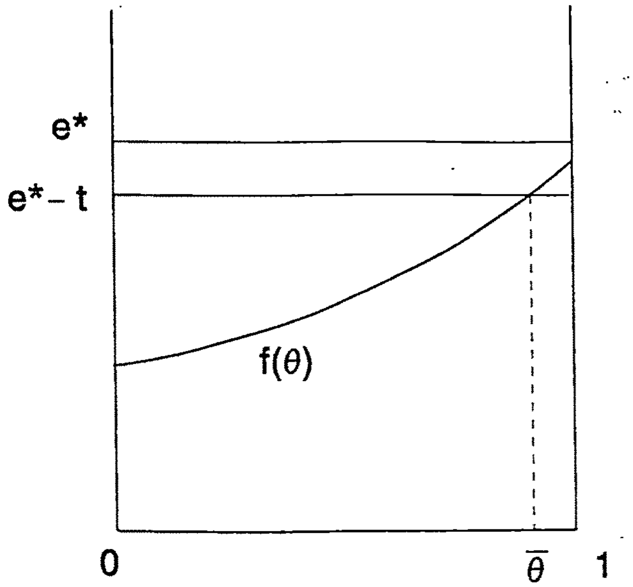


FIGURE 2. THE MANAGED EXCHANGE RATE AND THE EXCHANGE RATE IN THE ABSENCE OF INTERVENTION AS A FUNCTION OF THE STATE OF FUNDAMENTALS

off to abandoning the exchange rate is thus zero while the payoff to defending the exchange rate is

$$v - c(\alpha, \theta).$$

We assume that c is continuous and is increasing in α while decreasing in θ. In particular, to make the problem economically interesting we will impose the following assumptions on the cost function and the floating exchange rate $f(\theta)$.

- $c(0, 0) > v$. In the worst state of fundamentals, the cost of defending the currency is so high that it exceeds the value v even if no speculators attack.
- $c(1, 1) > v$. If all the speculators attack the currency, then even in the best state of the fundamentals, the cost of defending the currency exceeds the value.
- $e^* - f(1) < t$. In the best state of the fundamentals, the floating exchange rate $f(1)$ is sufficiently close to the pegged level e^* such that any profit from the depreciation of the currency is outweighed by the transactions cost t.

Let us denote by $\underline{\theta}$ the value of θ which solves $c(0, \theta) = v$. In other words, $\underline{\theta}$ is the value of θ at which the government is indifferent between defending the peg and abandoning it in the absence of any speculative selling. When $\theta < \underline{\theta}$, the cost of defending the currency exceeds the value, even if no speculators attack the currency (see Figure 1). At the other end, denote by $\overline{\theta}$ the value of θ at which $f(\theta) = e^* - t$, so that the floating exchange rate is below the peg by the amount of the cost of attack. When $\theta > \overline{\theta}$, then the floating exchange rate is sufficiently close to the peg that a speculator cannot obtain a positive payoff by attacking the currency (see Figure 2). Using the two benchmark levels of the state of fundamentals $\underline{\theta}$ and $\overline{\theta}$, we can classify the state of fundamentals under three headings, according to the underlying strategic situation.

A. *Tripartite Classification of Fundamentals*

Assuming that $\underline{\theta} < \overline{\theta}$, we can partition the space of fundamentals into three intervals,[3] emphasized by Maurice Obstfeld (1996).

- In the interval $[0, \underline{\theta}]$, the value of defending the peg is outweighed by its cost irrespective of the actions of the speculators. The government then has no reason to defend the currency. For this reason, we say that the currency is *unstable* if $\theta \in [0, \underline{\theta}]$.

---

[3] This assumption will hold if v is large and t is small.

- In the interval $(\underline{\theta}, \overline{\theta})$, the value of defending the currency is greater than the cost, provided that sufficiently few speculators attack the currency. In particular, if none of the speculators attacks, then the value of defending the currency is greater than the cost, and so the government will maintain the peg, which in turn justifies the decision not to attack. However, it is also the case that if all the speculators attack the currency, then the cost of defending the currency is too high, and the government will abandon the peg. Moreover, since $\overline{\theta}$ is the right end point of this interval, a speculator will make a positive profit if the government were to abandon the peg at any state θ in the interval $(\underline{\theta}, \overline{\theta})$, so that if a speculator believes that the currency peg will be abandoned, then attacking the currency is the rational action. For this reason, we say that the currency is *ripe for attack* if $\theta \in (\underline{\theta}, \overline{\theta})$.
- Finally, in the interval $[\overline{\theta}, 1]$, although the speculators can force the government to abandon the peg, the resulting depreciation of the currency is so small that they cannot recoup the cost of attacking the currency. Thus, even if a speculator were to believe that the currency will fall, the rational action is to refrain from attacking it. In other words, it is a dominant action not to attack. For this reason, we say that the currency is *stable* if $\theta \in [\overline{\theta}, 1]$.

The interesting range is the "ripe for attack" region. Suppose that the government's decision on whether or not to defend the currency is determined purely by weighing up the costs and benefits, and that it makes its decision once all the speculators have made their decisions. Then, if all the speculators have perfect information concerning the realization of θ, the "ripe for attack" region gives rise to the standard case of multiple equilibria due to the self-fulfilling nature of the speculators' beliefs. If the speculators believe that the currency peg will be maintained, then it is optimal not to attack, which in turn induces the government to defend the currency, thereby vindicating the speculators' decisions not attack the currency. On the other hand, if the speculators believe that the currency peg will be abandoned, the rational action is to attack the currency, which in turn induces the government to abandon the peg, vindicating the decision to attack. Given this multiplicity of equilibria, no definitive prediction can be made as to whether the currency will come under attack or not. We will now see, however, that the situation is very different when the speculators face a small amount of uncertainty concerning the fundamentals. Each state of fundamentals gives rise to a unique outcome.

B. *Game with Imperfect Information of Fundamentals*

Suppose that the speculators each have a signal concerning the state of fundamentals, as in the following game.

- Nature chooses the state of fundamentals according to the uniform density over the unit interval.
- When the true state is θ, a speculator observes a signal x which is drawn uniformly from the interval $[\theta - \varepsilon, \theta + \varepsilon]$, for some small[4] $\varepsilon > 0$. Conditional on θ, the signals are identical and independent across individuals. Based on the signal observed, a speculator decides whether or not to attack the currency.
- The government observes the realized proportion of speculators who attack the currency, α, and observes θ.

The payoffs of the game follow from the description of the model above. We assume that if a speculator is indifferent between attacking and not attacking, he will refrain from attacking and that if the government is indifferent between defending the peg and abandoning it, it will choose to abandon it.[5]

An equilibrium for this game consists of strategies for government and for the con-

---

[4] In particular, we assume that $2\varepsilon < \min\{\underline{\theta}, 1 -$

[5] Nothing substantial hinges on these assumptions, which are made for purposes of simplifying the statement of our results.

continuum of speculators such that no player has an incentive to deviate. We can solve out the government's strategy at the final stage of the game, to define a reduced-form game between the speculators only. To do this, consider the critical proportion of speculators needed to trigger the government to abandon the peg at state θ. Let $a(\theta)$ denote this critical mass. In the "unstable" region $a(\theta) = 0$, while elsewhere $a(\theta)$ is the value of α which solves $c(\alpha, \theta) = v$. Figure 3 depicts this function, which is continuous and strictly increasing in θ where it takes a positive value, and is bounded above by 1.

The unique optimal strategy for the government is then to abandon the exchange rate only if the observed fraction of deviators, α, is greater than or equal to the critical mass $a(\theta)$ in the prevailing state θ.

Taking as given this optimal strategy for the government, we can characterize the payoffs in the reduced-form game between the speculators. For a given profile of strategies of the speculators, we denote by

$$\pi(x)$$

the proportion of speculators who attack the currency when the value of the signal is x. We denote by $s(\theta, \pi)$ the proportion of speculators who end up attacking the currency when the state of fundamentals is θ, given aggregate selling strategy π. Since signals are uniformly distributed over $[\theta - \varepsilon, \theta + \varepsilon]$ at θ, we have[6]

$$(1)\quad s(\theta, \pi) = \frac{1}{2\varepsilon} \int_{\theta-\varepsilon}^{\theta+\varepsilon} \pi(x)\, dx.$$

Denote by $A(\pi)$ the event where the government abandons the currency peg if the speculators follow strategy π:

$$(2)\quad A(\pi) = \{\theta \mid s(\theta, \pi) \geq a(\theta)\}.$$

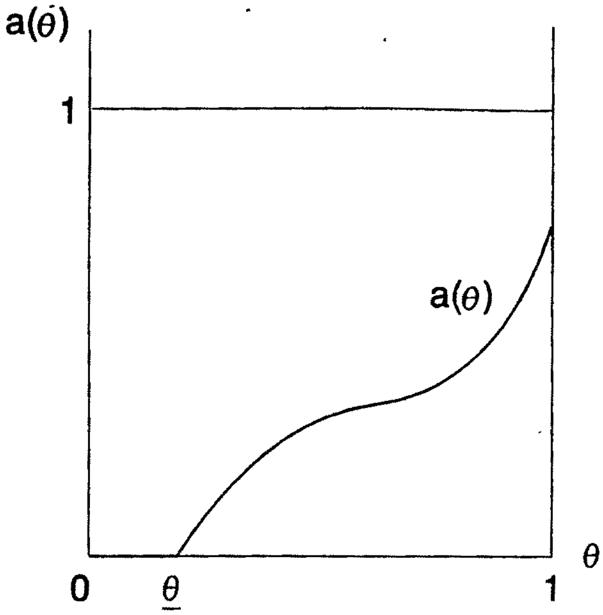


FIGURE 3. THE PROPORTION OF SPECULATORS WHOSE SHORT SALES ARE SUFFICIENT TO INDUCE DEPRECIATION, EXPRESSED AS A FUNCTION OF THE FUNDAMENTALS

We can then define the payoffs of a reduced-form game between the speculators. The payoff to a speculator of attacking the currency at state θ when aggregate short sales are given by π is

$$(3)\quad h(\theta, \pi) \equiv \begin{cases} e^* - f(\theta) - t & \text{if } \theta \in A(\pi) \\ -t & \text{if } \theta \notin A(\pi). \end{cases}$$

However, a speculator does not observe θ directly. The payoff to attacking the currency must be calculated from the posterior distribution over the states conditional on the signal x. The expected payoff to attacking the currency conditional on the signal x is given by the expectation of (3) conditional on x. Denoting this by $u(x, \pi)$, we have

$$(4)\quad u(x, \pi) = \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} h(\theta, \pi)\, d\theta = \frac{1}{2\varepsilon} \left[\int_{A(\pi) \cap [x-\varepsilon, x+\varepsilon]} (e^* - f(\theta))\, d\theta \right] - t.$$

---

[6] The following formula is for $\theta \in [\varepsilon, 1 - \varepsilon]$; for θ close to 0 or 1, the limits of the integration must be adjusted accordingly.

Since a speculator can guarantee a payoff of zero by refraining from attacking the currency, the rational decision conditional on signal x depends on whether $u(x, \pi)$ is positive or negative. Thus if the government follows its unique optimal strategy, π is an equilibrium of the first period game if $\pi(x) = 1$ whenever $u(x, \pi) > 0$ and $\pi(x) = 0$ whenever $u(x, \pi) \leq 0$.

II. Unique Equilibrium

We now state the main result of our paper, noting the contrast between the multiplicity of possible outcomes when there is perfect information of the fundamentals against the uniqueness of outcome when there is a small amount of noise.

THEOREM 1: *There is a unique θ^* such that, in any equilibrium of the game with imperfect information, the government abandons the currency peg if and only if $\theta \leq \theta^*$.*

The argument for our result can be presented in three steps.

LEMMA 1: *If $\pi(x) \geq \pi'(x)$ for all x, then $u(x, \pi) \geq u(x, \pi')$ for all x.*

In other words, if we compare two strategy profiles π and π', where π entails a greater proportion of speculators who attack for any message x, then the payoff to attacking the currency is greater given π than when it is given by π'. Thus speculators' decisions to attack the currency are strategic complements.

PROOF OF LEMMA 1:

Since $\pi(x) \geq \pi'(x)$, we have $s(\theta, \pi) \geq s(\theta, \pi')$ for every θ, from the definition of s given by (1). Thus, from (2),

$$A(\pi) \supseteq A(\pi').$$

In other words, the event in which the currency peg is abandoned is strictly larger under π. Then, from (4) and the fact that $e^* - f(\theta)$ is nonnegative,

$$\begin{aligned} & u(x, \pi) \\ & = \frac{1}{2\varepsilon}\left[\int_{A(\pi)\cap[x-\varepsilon, x+\varepsilon]} (e^* - f(\theta))\, d\theta\right] - t \\ & \geq \frac{1}{2\varepsilon}\left[\int_{A(\pi')\cap[x-\varepsilon, x+\varepsilon]} (e^* - f(\theta))\, d\theta\right] - t \\ & = u(x, \pi'), \end{aligned}$$

which proves the lemma.

For the next step, consider the strategy profile where every speculator attacks the currency if and only if the message x is less than some fixed number k. Then, aggregate short sales π will be given by the indicator function I_k, defined as

$$I_k(x) = \begin{cases} 1 & \text{if } x < k \\ 0 & \text{if } x \geq k \end{cases} \tag{5}$$

When speculators follow this simple rule of action, the expected payoff to attacking the currency satisfies the following property.

LEMMA 2: *$u(k, I_k)$ is continuous and strictly decreasing in k.*

In other words, when aggregate short sales are governed by I_k, and we consider the payoff to attacking the currency given the marginal message k, this payoff is decreasing as the fundamentals of the economy become stronger. Put another way, when the fundamentals of the economy are stronger, the payoff to attacking the currency is lower for a speculator on the margin of switching from attacking to not attacking. Such a property would be a reasonable feature of any model of currency attacks where the government is able to resist speculators better when the fundamentals are stronger.

The proof of Lemma 2 is simple but involves some algebraic manipulation, and hence is presented separately in the Appendix. Taking Lemma 2 as given, we can then prove the following result.

LEMMA 3: *There is a unique x^* such that, in any equilibrium of the game with imperfect* *information of the fundamentals, a speculator with signal x attacks the currency if and only if x < x*.*

To prove this, we begin by establishing that there is a unique value of k at which

$$u(k, I_k) = 0.$$

From Lemma 2, we know that $u(k, I_k)$ is continuous and strictly decreasing in k. If we can show that it is positive for small values of k and negative for large values, then we can guarantee that $u(k, I_k) = 0$ for some k. When k is sufficiently small (i.e., $k \leq \underline{\theta} - \varepsilon$), the marginal speculator with message k knows that the true state of fundamentals is in the "unstable" region, since such a message is consistent only with a realization of θ in the interval $[0, \underline{\theta}]$. Since the payoff to attacking the currency is positive at any θ in this interval, we have $u(k, I_k) > 0$. Similarly, when k is sufficiently large (i.e., $k \geq \overline{\theta} + \varepsilon$), the marginal speculator with message k knows that the true state of fundamentals is in the "stable" region. Since the payoff to attacking is negative at every state in this region, we have $u(k, I_k) < 0$. Hence, there is a unique value of k for which $u(k, I_k) = 0$, and we define the value x^* as this unique solution to $u(k, I_k) = 0$.

Now, consider any equilibrium of the game, and denote by $\pi(x)$ the proportion of speculators who attack the currency given message x. Define the numbers $\underline{x}$ and $\overline{x}$ as

$$\underline{x} = \inf\{\, x \mid \pi(x) < 1 \,\}$$

$$\text{and } \overline{x} = \sup\{\, x \mid \pi(x) > 0 \,\}.$$

Since $\overline{x} \geq \sup\{x \mid 0 < \pi(x) < 1\} \geq \inf\{x \mid 0 < \pi(x) < 1\} \geq \underline{x}$,

$$\underline{x} \leq \overline{x}. \tag{6}$$

When $\pi(x) < 1$, there are some speculators who are not attacking the currency. This is only consistent with equilibrium behavior if the payoff to not attacking is at least as high as the payoff to attacking given message x. By continuity, this is true at $\underline{x}$ also. In other words,

$$u(\underline{x}, \pi) \leq 0. \tag{7}$$

Now, consider the payoff $u(\underline{x}, I_{\underline{x}})$. Clearly, $I_{\underline{x}} \leq \pi$, so that Lemma 1 and (7) imply $u(\underline{x}, I_{\underline{x}}) \leq u(\underline{x}, \pi) \leq 0$. Thus, $u(\underline{x}, I_{\underline{x}}) \leq 0$. Since we know from Lemma 2 that $u(k, I_k)$ is decreasing in k and x^* is the unique value of k which solves $u(k, I_k) = 0$, we have

$$\underline{x} \geq x^*. \tag{8}$$

A symmetric argument establishes that

$$\overline{x} \leq x^*. \tag{9}$$

Thus, from (8) and (9), we have $\underline{x} \geq x^* \geq \overline{x}$. However, we know from (6) that this implies

$$\underline{x} = x^* = \overline{x}.$$

Thus, the equilibrium π is given by the step function I_{x^*}, which is what Lemma 3 states. Hence, this proves Lemma 3.

From this, it is a short step to the proof of our main theorem itself. Given that equilibrium π is given by the step function I_{x^*}, the aggregate short sales at the state θ are given by

$$s(\theta, I_{x^*}) = \begin{cases} 1 & \text{if } \theta < x^* - \varepsilon \\ \dfrac{1}{2} - \dfrac{1}{2\varepsilon}(\theta - x^*) & \\ \quad \text{if } x^* - \varepsilon \leq \theta < x^* + \varepsilon & \\ 0 & \text{if } \theta \geq x^* + \varepsilon. \end{cases}$$

Aggregate short sales $s(\theta, I_{x^*})$ are decreasing in θ when its value is strictly between 0 and 1, while $a(\theta)$ is increasing in θ for this range. See Figure 4, which illustrates the derivation of the cutoff point for the state of fundamentals at which the equilibrium short sales are equal to the short sales which induce depreciation.

We know that $x^* > \underline{\theta} - \varepsilon$, since otherwise attacking the currency is a strictly better action, contradicting the fact that x^* is a switching point. Thus, $s(\theta, I_{x^*})$ and $a(\theta)$ cross precisely once. Define θ^* to be the value of θ at which these two curves cross. Then, $s(\theta, I_{x^*}) \geq a(\theta)$ if and only if $\theta \leq \theta^*$, so that the government abandons the currency peg if and only if $\theta \leq \theta^*$. This is the claim of our main theorem.

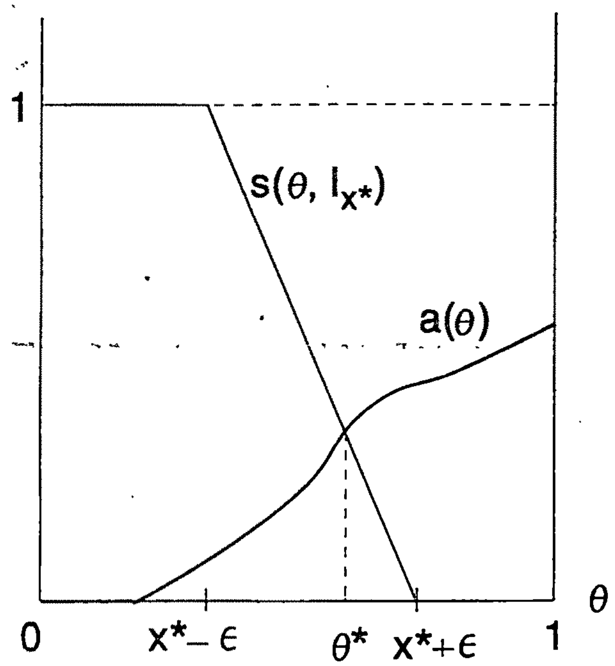


FIGURE 4. THE DERIVATION OF THE CUTOFF POINT FOR THE STATE OF FUNDAMENTALS AT WHICH THE EQUILIBRIUM SHORT SALES ARE EQUAL TO THE SHORT SALES WHICH INDUCE DEPRECIATION

III. Comparative Statics and Policy Implications

A. *Changes in the Information Structure*

When there is no noise, there are multiple equilibria throughout the "ripe for attack" region of fundamentals. But when there is positive noise, there is a unique equilibrium with critical value θ^*. The value of θ^* in the limit as ε tends to zero has a particularly simple characterization.

THEOREM 2: *In the limit as ε tends to zero, θ^* is given by the unique solution to the equation $f(\theta^*) = e^* - 2t$.*

The proof is in the Appendix. Intuition is gained by considering the marginal speculator who observes message $x = \theta^*$. With ε small, this tells the speculator that the true θ is close to θ^*. Since the government abandons the peg if and only if θ is less than θ^*, he attaches equal probability to the currency being abandoned and defended. So, the expected payoff to attacking is $\frac{1}{2}(e^* - f(\theta^*))$, while the cost is t. For the marginal speculator, these are equal, leading to the equation in Theorem 2.

Information plays a subtle role in the model. What matters is not the amount of information per se, but whether there is common knowledge. With noise, it is never common knowledge that the fundamentals are consistent with the government maintaining the currency peg, i.e., that $\theta \geq \underline{\theta}$. If you observe a signal greater than $\underline{\theta} + \varepsilon$, then you can claim to *know* that $\theta \geq \underline{\theta}$, since your message has the margin of error of ε. When do you know that *everyone* knows that $\theta \geq \underline{\theta}$? In other words, when do you know that everyone has observed a signal greater than $\underline{\theta} + \varepsilon$? Since others' signals can differ from yours by at most 2ε, this will be true if you observe a signal greater than $\underline{\theta}$ + 3ε. Proceeding in this way, there is "nth order knowledge" that $\theta \geq \underline{\theta}$ (i.e., everyone knows that everyone knows ... (n times) that everyone knows it) exactly if everyone has observed a signal greater than or equal to $\underline{\theta}$ + $(2n - 1)\varepsilon$. But, by definition, there is common knowledge that $\theta \geq \underline{\theta}$ if and only if there is nth order knowledge for every n. But for any fixed ε and signal x, there will be some level n at which nth order iterated knowledge fails. Thus it is never common knowledge that θ is not in the unstable region.

One interpretation we may put on noisy information is that the recipients of the differential information learn of the true underlying fundamentals of the economy with little error, but that there are small discrepancies in how these messages are interpreted by the recipients. When looking for a cause or a trigger for a currency attack, we should look for the arrival of noisy information, i.e., news events that are not interpreted in exactly the same way by different speculators. The information events that matter may be quite subtle. A "grain of doubt," allowing that others may believe that the economy is, in fact, unstable, will lead to a currency crises even if everyone knows that the economy is not unstable. In predicting when crises will occur, average opinion or even extreme opinion need not precipitate a crisis. Rather, what matters is the higher order beliefs of some participants who are apprehensive about the beliefs of others concerning the beliefs of yet further individuals, on these extreme opinions. This interpretation may shed light on some accounts of recent currency crises. Rumors of politi trouble in Chiapas Province were widely cited as a cause of the 1994 Mexico crisis (*New York Times*, 1994); uncertainty about Maastricht, German unification, and Bundesbank pronouncements were argued to be important in the 1992 European Monetary System crises. Our analysis suggests that such "informational events" might precipitate a crisis even if no investor thought they conveyed real information about fundamentals themselves. It is enough that the announcements remove common knowledge that the fundamentals were sustainable.

Above all, our analysis suggests an important role for public announcements by the monetary authorities, and more generally, the transparency of the conduct of monetary policy and its dissemination to the public. If it is the case that the onset of currency crises may be precipitated by higher-order beliefs, even though participants believe that the fundamentals are sound, then the policy instruments which will stabilize the market are those which aim to restore transparency to the situation, in an attempt to restore common knowledge of the fundamentals. The most effective means towards this would be a prominent public announcement which is commonly known to convey information to all relevant participants. The canonical case of a communication arrangement which would be conducive to achieving common knowledge is the "town hall meeting" in which an announcement is made to an audience gathered in a single room, where everyone can observe that all other participants are in an identical position. In contrast, if the audience is fragmented, and must communicate in small groups, common knowledge is extremely difficult to achieve.[7] The Clinton administration's announcement of the $40 billion dollar rescue package for the Mexican peso can be seen as an attempt to restore the sort of transparency referred to above. Its effectiveness derived more from its very public nature, rather than the actual sum of money involved. This suggests a crucial role for the timely and effective dissemination of information on the part of policy makers, and the smooth functioning of a reliable and transparent set of communication channels between market participants and the policy makers, as well as between the market participants themselves.

B. *Changes in Transaction Costs*

We next examine shifts in t. Drawing on Theorem 2, which determines θ^* in the limit as ε becomes small, we can totally differentiate the equation $f(\theta^*) = e^* - 2t$ to obtain

$$\frac{d\theta^*}{dt} = -\frac{2}{f'(\theta^*)}.$$

Thus increasing transaction costs prevents currency crises, since it reduces the range of fundamentals where an attack occurs. The size of this effect depends on the slope of f: when the f' is small, an increase in the cost of speculation has a large effect on the switching point θ^*. This suggests that the imposition of small transactions costs as advocated by several commentators will have a large impact on the prevalence of speculation precisely when the *consequences* of such speculative attacks are small. If speculative attacks are predicted to lead to drastic effects (i.e., when $f(\theta)$ is a steep function of θ), then the imposition of a small additional cost is unlikely to have a large effect on the incidence of currency attacks.

C. *Changes in Aggregate Wealth*

Now consider how our analysis is affected when aggregate wealth varies. The international flow of so-called "hot money" would be one factor in determining this aggregate wealth, as well as changes in the numbers of speculators themselves. The main effect of a change in aggregate wealth of the speculators is a change in the function $a(\theta)$, which indicates the critical proportion of speculators needed to attack the currency in order to induce the government to abandon the currency peg. When the aggregate wealth of the speculators increases, then this critical proportion of speculators falls, since the government's decision is based on the absolute *level* of short sales.

---

[7] Ariel Rubinstein's (1989) e-mail game is a case in point, and Michael S.-Y. Chwe (1996) has suggested some of the relevant factors which would allow us to address this issue in a more general context.

As can be seen from Figure 4, a downward shift in the $a(\theta)$ function has the effect of enlarging the set of states at which the government abandons the exchange rate peg. In other words the event $A(\pi) = \{\theta \mid s(\theta, \pi) > a(\theta)\}$ is strictly larger with a lower $a(\cdot)$ function. Since the payoff to speculation is given by

$$\int_{A(\pi)\cap[x-\varepsilon,x+\varepsilon]} (e^* - f(\theta))\, d\theta - t,$$

the enlargement of the event $A(\pi)$ has an unambiguous effect in increasing the payoff to attacking the currency at any value of the signal x. Thus, the benchmark value θ^* is shifted to the right, and the incidence of speculative attacks increases.

Note, however, from Figure 4 that the effect of an increase in the $a(\cdot)$ function depends on the size of the noise ε. The effect is largest when ε is also large. In the limiting case when ε tends to zero, the equilibrium $s(\theta, I_{x^*})$ becomes the step function I_{x^*}, so that a shift in the $a(\cdot)$ function has no effect on the switching point θ^*. Thus, our analysis suggests that changes in the aggregate wealth of speculators need not have a large impact on the incidence of currency attacks when the speculators have fairly precise information concerning the fundamentals. It is when the noise is large that shifts in wealth have a big impact.

This suggests that the imposition of direct capital controls work best when there is a lack of "transparency" of the economic fundamentals, in the sense that observers differ widely in their interpretation of the economic fundamentals. When the fundamentals are relatively transparent to all (corresponding to a small ε), direct capital controls seem far less effective. Under such circumstances, strategic considerations dominate any uncertainty concerning the fundamentals.

IV. Conclusion

Existing models of currency attacks that focus on fundamentals ignore the role of speculators' beliefs about other speculators' behavior. Existing self-fulfilling beliefs models of currency attacks assume that speculators *know* (in equilibrium) exactly what other speculators will do. Neither feature is realistic. Our model takes neither extreme. Because there some uncertainty about equilibrium, speculators are uncertain as to exactly what other speculators will do; but their behavior depends nontrivially on what they believe they will do. Because our model of self-fulfilling currency attacks is consistent with unique equilibrium, we are able to analyze the impact of alternative policies.

APPENDIX

PROOF OF LEMMA 2:

Consider the function $s(\theta, I_k)$, which gives the proportion of speculators who attack the currency at θ when the aggregate short sales is given by the step function I_k. Since x is uniformly distributed over $[\theta - \varepsilon, \theta + \varepsilon]$, we have

$$\text{(A1)}\quad s(\theta, I_k) = \begin{cases} 1 & \text{if } \theta \le k - \varepsilon \\ \dfrac{1}{2} - \dfrac{1}{2\varepsilon}(\theta - k) & \text{if } k - \varepsilon \le \theta \le k + \varepsilon \\ 0 & \text{if } \theta \ge k + \varepsilon. \end{cases}$$

If aggregate short sales are given by I_k, there is a unique θ (which depends on k) where the mass of speculators attacking equals the mass of speculators necessary to cause the government to abandon the exchange rate [where $s(\theta, I_k) = a(\theta)$]. Write $\psi(k)$ for the amount that θ must exceed k in order for this to be true. In other words, $\psi(k)$ is the unique value of ψ solving $s(k + \psi, I_k) = a(k + \psi)$. Observe that $\psi(k) = \varepsilon$ if $k \le \underline{\theta} - \varepsilon$, while if $k > \underline{\theta} - \varepsilon$, then $-\varepsilon < \psi(k) < \varepsilon$ and is the value of ψ solving $(1/2 - \psi/2\varepsilon) = a(k + \psi)$.

Since the government abandons the currency peg if and only if θ lies in the interval $[0, k + \psi(k)]$, the payoff function $u(k, I_k)$ is given by

$$\text{(A2)}\quad \frac{1}{2\varepsilon}\int_{k-\varepsilon}^{k+\psi(k)} (e^* - f(\theta))\, d\theta - t.$$

Since $e^* - f(\theta)$ is strictly decreasing in θ, if we can show that $\psi(k)$ is weakly decreasing in k, this will be sufficient to show that $u(k, I_k)$ is strictly decreasing in k. To see that is weakly decreasing in k, totally differentiate the equation $(1/2 - \psi/2\varepsilon) = a(k + \psi)$ with respect to k, to obtain $-(1/2\varepsilon)\psi'(k) = a'(\theta)(1 + \psi'(k))$. Hence,

$$\psi'(k) = -\frac{a'(\theta)}{a'(\theta) + (1/2\varepsilon)} \leq 0,$$

which is sufficient for $u(k, I_k)$ to be strictly decreasing in k. Finally, the continuity of $u(k, I_k)$ follows immediately from the fact that it is an integral in which the limits of integration are themselves continuous in k. This completes the proof of Lemma 2.

PROOF OF THEOREM 2:

Consider the switching point x^*, which is the solution to the equation $u(x^*, I_{x^*}) = 0$. Then, writing $F(\varepsilon) = \int_{A(I_{x^*}) \cap [x^* - \varepsilon, x^* + \varepsilon]} (e^* - f(\theta))\, d\theta$, we can express this equation as $(F(\varepsilon)/2\varepsilon) - t = 0$. By using L'Hôpital's rule,

$$\lim_{\varepsilon \to 0} \frac{F(\varepsilon)}{2\varepsilon} = \frac{F'(0)}{2} = \frac{e^* - f(x^*)}{2}.$$

Thus, in the limit as ε → 0, equation (A3) yields $f(x^*) = e^* - 2t$. Finally, we note that x^* converges to θ^* when ε tends to zero, since in the limit, $s(\theta, I_{x^*}) = I_{x^*}$.

REFERENCES

Calvo, Guillermo A. "Servicing the Public Debt: The Role of Expectations." *American Economic Review*, September 1988, *78*(4), pp. 647–61.

Carlsson, Hans and van Damme, Eric E. "Global Games and Equilibrium Selection." *Econometrica*, September 1993, *61*(5), pp. 989–1018.

Chwe, Michael S.-Y. "Structure and Strategy in Collective Action: Communication and Coordination in Social Networks." Mimeo, University of Chicago, 1996.

Cole, Harold L. and Kehoe, Timothy J. "Self-Fulfilling Debt Crises." Federal Reserve Bank of Minneapolis Research Department Staff Report No. 211, 1996.

Dornbusch, Rudiger and Werner, Alejandro M. "Mexico: Stabilization, Reform and No Growth." *Brookings Papers on Economic Activity*, 1994, (1), pp. 253–315.

Eaton, Jonathan and Fernandez, Raquel. "Sovereign Debt," in Gene M. Grossman and Kenneth Rogoff, eds., *Handbook of international economics*, Vol. III. Amsterdam: North-Holland, 1995, pp. 2031–71.

Eichengreen, Barry; Tobin, James and Wyplosz, Charles. "Two Cases for Sand in the Wheels of International Finance." *Economic Journal*, January 1995, *105*(428), pp. 162–72.

Eichengreen, Barry and Wyplosz, Charles. "The Unstable EMS." *Brookings Papers on Economic Activity*, 1993, (1), pp. 51–143.

Flood, Robert P. and Garber, Peter M. "Gold Monetization and Gold Discipline." *Journal of Political Economy*, February 1984, *92*(1), pp. 90–107.

Morris, Stephen; Rob, Rafael and Shin, Hyun Song. "*p*-Dominance and Belief Potential." *Econometrica*, January 1995, *63*(1), pp. 145–57.

New York Times, The. December 20–23, 1994, various articles.

Obstfeld, Maurice. "Rational and Self-Fulfilling Balance-of-Payments Crises." *American Economic Review*, March 1986, *76*(1), pp. 72–81.

______. "The Logic of Currency Crises." *Cahiers Economiques et Monetaires (Banque de France)*, 1994, *43*, pp. 189–213.

______. "Models of Currency Crises with Self-Fulfilling Features." *European Economic Review*, April 1996, *40*(3–5), pp. 1037–47.

Postlewaite, Andrew and Vives, Xavier. "Bank Runs as an Equilibrium Phenomenon." *Journal of Political Economy*, June 1987, *95*(3), pp. 485–91.

Rubinstein, Ariel. "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge.' " *American Economic Review*, June 1989, *79*(3), pp. 385–91.

Compensating Differentials for Gender-Specific Job Injury Risks

By JONI HERSCH*

The theory of compensating differentials holds a prominent place in the analysis of wage determination, and a large empirical literature documents substantial wage-risk trade-offs.[1] Although women comprise over 45 percent of the labor force, job risks have been viewed as primarily a male province, and studies have largely excluded female workers from the analyses. Since most of the occupational injury data are available only at the industry level, if women are in safer jobs within industries, matching industry injury rates to female workers may lead to large measurement problems and misleading estimates.

In this paper I use new national data to construct gender-specific estimates of injury and illness incidence rates by both industry and occupation. These rates are the first gender-specific injury incidence rates in the literature, as well as the first occupational incidence rates. Although women are less likely than men to be injured on their jobs, their injury experience is considerable, as one-third of all injuries and illnesses with days away from work are to women. Adjusted for gender differences in employment, women face a job risk that is 71 percent of men's. Further, the gender-specific injury rates reveal substantial and statistically significant wage-risk trade-offs for female workers. All female workers, not only those employed in blue-collar occupations, receive a significant compensating differential for job risks. Female workers at the average level of risk receive a wage premium of 2–3 percent, equal to an additional $400–$563 per year as compensation for nonfatal job risks.

The results using the gender-specific injury measures are in stark contrast to those obtained for women using the standard industry risk measures. Estimates based on the U.S. Department of Labor, Bureau of Labor Statistics (BLS), industry injury and illness incidence rates do not indicate a significant wage-risk trade-off. These findings suggest that assigning industry risk measures to female workers without adjusting for gender differences in injury experience may lead to biased estimates of the returns to job risk and a misleading view of who bears injury risks in the workplace.

I use these estimates of the wage-risk trade-off based on the gender-specific incidence rates to calculate the first estimates in the literature of the implicit value of an injury or illness for women workers. The values, which range from $20,000 using the female-specific industry rate, to $30,000 using the female-specific occupation rate, are similar to those I find for male blue-collar workers.

Similar concerns about measurement error have led to the exclusion of white-collar male workers from most studies. The industry- and occupation-specific risk measures calculated here certainly reduce this measurement error. In contrast to the findings for female workers, however, in many cases there is an inverse relation between wages and risk for white-collar males, whether the risk measure pertains to the individual's three-digit industry or three-digit occupation.

The estimating procedure used in this paper follows the standard approach in the literature, matching average risk measures to individuals by industry or occupation. Because all workers within an industry or occupation are assigned the same injury rate, the residuals in the regression for workers in a given industry or occupation group are likely to be correlated. The robust standard errors calculated in this paper are generally 2–3 times the size of those calculated without recognizing this source of correlation.

---

* Department of Economics and Finance, University of Wyoming, Laramie, WY 82071. I thank Kathryn Anderson, Stephen Cosslett, Jahn Hakes, Daniel Hamermesh, Leslie Stratton, and the anonymous referee for their very helpful suggestions, and Linda Garris and Larry Jones of the Office of Safety, Health, and Working Conditions, Bureau of Labor Statistics, for providing tables of injury statistics prior to their public release.

[1] A recent survey of the literature is by W. Kip Viscusi, 1993.

I. Industry Job Risks and the Treatment of Female Workers in the Literature

Before discussing the construction of gender-specific injury incidence rates, it is useful to consider the method used by the Bureau of Labor Statistics. For each industry, the incidence rate is calculated as

$$\text{(1) BLS Industry Rate} = (N/H) \times 200{,}000,$$

where N = number of injuries and illnesses, H = total hours worked by all employees during the calendar year, and 200,000 = base for 100 full-time equivalent workers (40 hours per week, 50 weeks per year.) These values are reported annually in the *Survey of Occupational Injuries and Illnesses in the United States, 1993* (BLS, 1995a).

These incidence rates pertain to all workers within an industry, so that, for instance, secretaries and miners within the mining industry are assigned the same risk measure. Since data on occupational injuries and on the gender distribution of injury cases were not available until recently, the standard practice in the literature has been to impute these industry risk values to all individuals in the wage sample by three-digit industry code. If workers with certain characteristics are in riskier or less risky jobs within their industry, however, the estimated returns to risk may be biased.[2] In order to reduce this potential source of bias, many authors limit their sample to male blue-collar workers or to workers in manufacturing. In addition, other restrictions are typically imposed, such as restrictions to household heads, hourly workers, or to full-time employees. Even among samples which do not explicitly exclude women, such restrictions severely limit the number of women eligible for inclusion in the analyses.

Given the small samples of female workers remaining after such restrictions have been imposed, most authors either exclude women entirely from the analysis, or include them by allowing gender to affect wages only through an intercept. While these approaches are reasonable in the absence of data on gender-specific injury experience, they do not allow tests of whether women receive a compensating differential for the job risks they face. Furthermore, since women are disproportionately employed in white-collar occupations, estimates of compensating differentials for blue-collar workers would not be representative of the population of women workers overall.

II. Gender-Specific Injury and Illness Incidence Rates

In this section, I calculate gender-specific injury and illness incidence rates for both industry and occupation. Recently, the BLS began collecting more extensive information on the worker and case characteristics of injury and illness cases involving days away from work. The restriction of the survey coverage to injuries and illnesses involving at least one day away from work provides a lower bound on the severity of the incidence, and increases the homogeneity of the definition of injury and

---

[2] In the hedonic wage model, an individual worker i is compensated for market beliefs about the objective riskiness R_i^* of his or her job. The wage equation for the ith individual can be written as

$$y_i = \beta R_i^* + \mu_i, \tag{i}$$

where y_i is the log of wage and μ_i is a random error term. In general, R_i^* is not observed, but is related to the observed average industry risk R_i as

$$R_i^* = R_i + \varepsilon_i, \tag{ii}$$

where ε_i is an unobserved risk component associated with the specific type of job held by the individual within the industry. The wage equation to be estimated is

$$y_i = \beta R_i + v_i, \tag{iii}$$

where $v_i = \beta\varepsilon_i + \mu_i$. If ε_i and μ_i are uncorrelated with R_i, then by solving for the OLS estimator of β and taking the expected value, one can show that the OLS estimator of β is unbiased. The variance of the error term in equation (iii) is larger than in equation (i), so the estimated standard errors of the coefficients in the wage regression will tend to be larger using group means. If ε_i is correlated with R_i (for instance if women tend to be in less risky jobs within industries), then using industry average risk can lead to biased and inconsistent estimates. This is a kind of omitted variable bias and can be either positive or negative.

illness.[3] These data reveal that female workers experience a surprising number of job injuries and illnesses. Of 2.25 million BLS-reported nonfatal occupational injuries and illnesses with days away from work in 1993, one-third occurred to women. It is notable that, at least in terms of duration, men and women suffer injuries of similar severity. The median days away from work for those experiencing such an injury or illness is five days for both male and female workers (BLS, 1995b). Furthermore, the same five injury types account for 68–69 percent of the injuries for both male and female workers.[4]

Although women's share of injuries of 32.7 percent is less than their employment share of 46 percent of private industry employees, the magnitude of their injury experience is quite surprising since women are largely concentrated in the safer white-collar occupations. Taking into account the different levels of employment by gender, women face a job risk that is 71 percent of men's. Only 20.8 percent of the cases with days away from work in 1993 occurred in white-collar occupations. Among private employees, 69 percent of the women, but only 43.5 percent of the men, are employed in white-collar occupations. Within white-collar occupations, the injury rate for women is 80 percent higher than for men.

To calculate gender-specific industry job risk rates, I use data from two BLS tables. These tables provide information on the number of cases with days away from work in 1993 by gender for three-digit or four-digit SIC code or three-digit occupation. The BLS does not calculate gender-specific incidence rates for either industry or occupation. Because of differences in the information available, I use different procedures, described below, to calculate these rates.

A. *Industry Incidence Rates*

In principle one could use equation (1) to calculate gender-specific incidence rates for each industry by replacing N and H with the corresponding gender-specific values. However, while the number of injuries and illnesses by gender are provided in the new BLS survey, total hours worked by industry and gender are not available. I therefore allocate the BLS average industry rate into gender-specific shares by weighing the BLS rate by the gender-specific share of cases relative to the gender-specific hours share for each industry i as follows:

$$(2)\quad \text{Gender Industry Rate} = (N_g/N)/(H_g/H) \times \text{BLS Industry Rate},$$

where N_g = total number of cases of gender g in industry i, and H_g = total hours worked by gender g in industry i. As in equation (1), N and H represent the total number of injury and illness cases and total hours worked by all employees.

To calculate gender-specific shares of total employment hours within three-digit industries, I use data from the *Census of Population and Housing, 1990* (U.S. Department of Commerce, Bureau of the Census, [1993]) 5 percent sample. Since government and self-employed workers are excluded from the BLS survey used to estimate injury incidence rates, I restrict the Census sample to paid employees in private industry who report working positive hours in the preceding week. This yields a sample of 4,149,478 observations.

Table 1 provides a comparison of the BLS industry incidence rate and the gender-specific rates for the major industry categories. For workers in private industry overall, the adjusted female incidence rate is 2.2 injury and illness cases with days away from work per 100 workers, considerably lower than the BLS average incidence rate of 2.9. Industries with larger shares of female employees, such as finance, insurance and real estate, and services have lower than average risk. As the fe[illegible] male incidence ratio in the last column

---

[3] Injury and illness cases not involving days away from work are far more common. There were about 67 percent more cases without days away from work in 1993 than with days away from work.

[4] These injuries are sprains, strains and tears; bruises and contusions; fractures; cuts and lacerations; and soreness, pain, hurt, except the back.

TABLE 1—INJURY AND ILLNESS INCIDENCE RATES WITH DAYS AWAY FROM WORK BY MAJOR INDUSTRY, 1993[a]

Industry	Percent female in industry[b]	BLS rate[c]	Female rate[d]	Male rate[d]	Female/male ratio
Private industry	45.9	2.9	2.2	3.4	0.65
Goods-producing					
Agriculture, forestry, and fishing	22.9	4.2	3.5	4.6	0.76
Mining	13.5	3.3	0.8	4.1	0.19
Construction	10.4	4.9	1.1	5.3	0.21
Manufacturing	33.6	3.3	2.5	3.4	0.75
Service-producing					
Transportation and public utilities	30.4	4.3	2.5	4.6	0.55
Wholesale trade	31.0	2.8	1.3	3.7	0.35
Retail trade	53.1	2.7	2.3	3.1	0.74
Finance, insurance, and real estate	63.5	1.0	0.9	1.3	0.65
Services	64.7	2.3	2.4	2.4	1.00

[a] Per 100 full-time workers.
[b] Author's calculation from *Census of Population and Housing, 1990.* Sample restricted to private, paid employees employed in industries reporting injury and illness cases by gender.
[c] *Survey of Occupational Injuries and Illnesses, 1993.*
[d] Author's calculations. See text.

cases; on average women face considerably less risk than men in the high-risk industries such as mining and construction that employ relatively few women. This indicates that there is considerable occupational sorting by gender within these industries.

The correlation between the three-digit BLS industry rate and the female-specific industry rate is 0.67. The corresponding correlation between the BLS rate and the male-specific rate is 0.96. This suggests that estimates of wage-risk trade-offs for men are likely to be similar using either the BLS industry rate or the gender-specific rate, but this is less likely to be true for women.

B. *Occupational Incidence Rates*

The BLS does not provide occupational injury and illness incidence rates, so the procedure used to estimate industry incidence rates cannot be used. A modification of the BLS equation (1) leads to estimates of occupational risk for each occupation k:

(3) Gender Occupation Rate

$$= (O_g / H_{go}) \times 200{,}000,$$

where O_g = number of cases for gender g in occupation k, and H_{go} = total hours worked by gender g in occupation k. Since the BLS does not provide occupational employment values, I again use Census data to estimate employment within each occupation.

Table 2 lists representative incidence rates for occupations with a large number of cases for female workers. For comparison, the corresponding incidence rates for men in these occupations are also included. While women generally face less risk than men within occupations, in most the gap is fairly narrow.

TABLE 2—SELECTED INJURY AND ILLNESS INCIDENCE RATES BY OCCUPATION, 1993

Occupation title	Percent female hours in occupation	Female rate	Male ra
Secretaries	98.9	0.46	1.03
Bank tellers	91.2	0.83	0.37
Cashiers	78.5	1.50	1.77
Registered nurses	94.5	2.12	3.39
Health aides, except nursing	81.2	7.80	9.04
Nursing aides, orderlies, and attendants	89.2	8.05	7.85
Miscellaneous food preparation	45.8	8.33	8.09
Truck drivers	4.3	9.63	6.62
Laborers, except construction	21.5	9.71	15.74
Public transportation attendants	80.0	11.14	7.76

Source: Author's calculations. See text.

Furthermore, within many occupations, such as truck drivers and public transportation attendants, women actually face greater risk than men.

III. Empirical Specification and Data

In order to test for the presence of compensating differentials, I estimate wage equations for both female and male workers of the following form:

$$(4)\qquad \ln(WAGE_i) = \alpha + \beta\, RISK_i + \Sigma_j \gamma_j \mathbf{X}_{ij} + \varepsilon_i,$$

where *WAGE* is the hourly wage rate; *RISK* is a measure of job risk; **X** is a vector of explanatory variables such as years of work experience, education, union status, and occupation; α, β, and γ_j are parameters to be estimated; and ε is a random error term. The prediction of hedonic wage theory is that $\beta > 0$.

Equation (4) is the standard specification used in the hedonic wage literature. Since the ε_i may have different variances in different industries or occupations, many authors correct the standard errors for group heteroskedasticity. For comparability to the literature, I present these standard errors in Table 3 in parentheses.

However, since individuals within the same industry or occupation group are assigned the same risk rate, the residuals in the regression for workers in a given industry or occupation may be correlated. Standard errors not corrected for this correlation may be too small. I therefore use a procedure for robust estimation of the standard errors, which accounts for the within-group correlation in industry or occupation.[5] I present these in brackets below the standard errors corrected for group heteroskedasticity in Table 3.

To estimate the wage equations, I use data from the 1994 *Current Population Survey*

---

[5] Peter J. Huber (1967) and William H. Rogers (1993). The estimations are performed using *Stata Release 5.0* (StatCorp, 1997).

(CPS), U.S. Department of Commerce, Bureau of the Census. The wage equations include workers aged 18–65 whose hourly wage rate exceeds $2, and who provide complete information on all variables used in the analysis. Further restrictions corresponding to those made by the BLS in the scope of its job injury data collection are necessary in order to assign risk measures to the individuals in the study; that is, I exclude workers in public administration, self-employed workers, and private household workers. I also exclude workers employed in the agriculture, forestry, and fisheries industries. The resulting samples consist of 6,037 female and 5,960 male workers.

Hourly wage is the reported hourly wage for [illegible] percent of the women and 58 percent of the men, and is calculated from weekly pay and hours usually worked on this job for the remainder. Since information on actual work history is unavailable, I use years of potential experience, measured as age − education − 6. While this approximation is adequate for the purpose of this paper, comparisons by gender might lead to misleading conclusions, since potential experience overstates actual experience by a greater magnitude for women than for men.

Other variables in the wage equation include years of completed schooling and indicators of race and union status. Differences in cost of living that may affect wages are controlled for by indicators of region and city size. Industry and occupation characteristics other than job risk also have a direct effect on wage levels. To the extent that unobserved industry and occupation characteristics are correlated with job risk, the estimated returns to job risk may be biased. To reduce the likelihood of this source of bias, I include indicators of major occupation and industry categories.

Means and standard deviations for the risk measures are provided in the first column of Table 3. The BLS industry rate faced by women is lower than for private industry overall, and reflects the fact that women sort into safer industries. That women sort into safer jobs within industries in addition to sorting into safer industries can be seen by comparing the average BLS rate for women in the sample to the average female-specific rate, which is 16 percent lower than the average BLS rate.

IV. Wage Equation Estimates

A. *Female Workers*

Table 3, Panel A, presents coefficient estimates of the risk measures in the wage equations.[6] Equation (1) uses the customary BLS industry risk measure. Based on this measure, there is no evidence of a compensating differential for job risk for women workers. However, the estimates based on the gender-specific risk measures show strong evidence of compensating differentials. The results in column (2) using the female-specific industry risk measure indicates a wage-risk trade-off which is significant at the 1-percent level based on the heteroskedasticity-corrected standard errors, although it is no longer significant using the robust standard errors allowing for within-group correlation. However, the estimates in column (3) using the female-specific occupation risk measure reveals substantial and statistically significant effects (at the 5-percent level or better in 1-sided tests) based on either standard error. The magnitude of the estimated wage-risk trade-off is larger using the occupation risk measure than using the industry risk measure (0.014 and 0.009, respectively).

The results based on gender-specific injury and illness incidence rates strongly indicate that women do receive a compensating differential for their exposure to job risk. To determine whether the source of the job risk derives from the pervading riskiness of the industry, from the riskiness of the worker's specific job, or from both industry and occupation risk characteristics, I estimate wage equations including both industry and occupation risk. For instance, although secretarial jobs are quite safe, secretaries employed in textile mills may be exposed to various job hazards, such as cotton dust, that secretaries in insurance companies do not face.

Column (4) presents the estimates including the BLS rate as well as the occupational rate, followed by the estimates based on the female-specific industry and occupation rates in column

---

[6] Selectivity-corrected estimates of the wage equation for female workers are virtually identical and are available upon request.

TABLE 3—WAGE EQUATION ESTIMATES[a]

	Mean (standard deviation)	Dependent variable: log of hourly wage[b] (1)	(2)	(3)	(4)	(5)
Panel A: Female Workers (sample size = 6,037)						
BLS industry rate	2.48	0.443	—	—	−0.025	—
	(1.57)	(0.360)			(0.388)	
	—	[0.927]			[0.608]	
Female industry rate	2.14	—	0.897*	—	—	[illegible]
	(1.56)		(0.364)			[illegible]
	—		[0.953]			[illegible]
Female occupation rate	1.94	—	—	1.362*	1.369*	[illegible]
	(2.22)			(0.374)	(0.401)	[illegible]
	—			[0.740]	[0.739]	[illegible]
Hourly wage	10.60	—	—	—	—	—
	(7.39)					
Adjusted R^2	—	0.45	0.45	0.45	0.45	0.45
Panel B: Male Workers (sample size = 5,960)						
BLS industry rate	3.05	−1.959*	—	—	−1.894*	—
	(1.67)	(0.395)			(0.397)	
	—	[0.646]			[0.643]	
Male industry rate	3.35	—	−2.042*	—	—	[illegible]
	(1.80)		(0.366)			[illegible]
	—		[0.573]			[illegible]
Male occupation rate	3.23	—	—	−0.550	−0.465	[illegible]
	(3.58)			(0.262)	(0.260)	[illegible]
	—			[0.377]	[0.251]	[illegible]
Hourly wage	13.52	—	—	—	—	—
	(8.80)					
Adjusted R^2	—	0.45	0.45	0.45	0.45	[illegible]
Panel C: Male Blue-Collar Workers (sample size = 3,197)						
BLS industry rate	3.62	1.360*	—	—	1.319*	
	(1.56)	(0.452)			(0.454)	
	—	[1.073]			[0.079]	
Male industry rate	3.93	—	1.280*	—	—	[illegible]
	(1.65)		(0.435)			[illegible]
	—		[1.032]			[illegible]
Male occupation rate	5.24	—	—	0.224	0.158	[illegible]
	(3.78)			(0.236)	(0.239)	[illegible]
	—			[0.297]	[0.272]	[illegible]
Hourly wage	10.84	—	—	—	—	
	(6.50)					
Adjusted R^2	—	0.34	0.34	0.34	0.34	[illegible]

TABLE 3—*Continued.*

	Mean (standard deviation)	Dependent variable: log of hourly wage[b] (1)	(2)	(3)	(4)	(5)
Panel D: Male Hourly Blue-Collar Workers (sample size = 2,578)						
BLS industry rate	3.62 (1.52) —	1.578* (0.503) [1.133]	—	—	1.470* (0.503) [1.126]	—
Male industry rate	3.92 (1.61) —	—	1.472* (0.479) [1.085]	—	—	1.349* (0.480) [1.082]
Male occupation rate	5.42 (3.89) —	—	—	0.494** (0.252) [0.279]	0.431** (0.254) [0.275]	0.418** (0.255) [0.275]
Hourly wage	10.58 (6.61)	—	—	—	—	—
Adjusted R^2	—	0.37	0.37	0.37	0.37	0.37

[a] All coefficients are multiplied by 100. Data set is March 1994 *Current Population Survey*. Additional variables in each equation are a constant, potential experience, potential experience squared, education, and indicator variables for [illegible], nonwhite, three regions, and five city sizes. The female equations and the equations for all male workers also include indicators for nine occupations and six industries. The blue-collar male equations also include indicator variables for two occupations and for manufacturing. See text for definitions of variables.

[b] Standard errors corrected for group heteroskedasticity in parentheses; standard errors corrected for within-group correlation in brackets.

* Indicates significance at the 1-percent level, and ** indicates significance at the 5-percent level (1-sided tests).

(5). As the results show, the coefficient of occupation risk is not affected by the inclusion of the industry rate. The industry rate is not significantly different from zero after controlling for occupation risk. Thus the source of the wage-risk trade-off for the female sample is predominantly due to the riskiness of the occupation.

B. *Male Workers*

Panels B, C, and D of Table 3 summarize the risk coefficients from the corresponding equations for men. Estimates pooling white- and blue-collar men are reported in Panel B and indicate a significantly negative wage-risk relation using either the BLS industry risk measure, the gender-specific industry risk measure, or the gender-specific occupation measure. Since the occupational risk measures calculated here should circumvent the large measurement error that may result from assigning industry average risk measures to men in white-collar occupations, the negative wage-risk trade-off found for the full male sample is puzzling. It is possible that this results from pooling workers paid hourly with those on salary. For instance, salaried workers in risky jobs may be compensated by increased opportunities for promotion rather than directly for the riskiness of their jobs. However, the results restricted to hourly workers yield significantly negative effects of industry risk of about half the magnitude of that found for the full sample, while there is no significant effect of occupation risk. The results restricted to all white-collar workers are similar to those found for the full sample, with significant negative returns to both industry and occupation risk. Estimates restricted to white-collar males paid hourly indicate no significant wage-risk trade-off using any measure of risk.[7]

---

[7] These results are available upon request.

The results reported in Panel C, after making the customary restriction to blue-collar men, indicate the customary findings of a positive wage-risk trade-off using both measures of industry risk, which are significant based on the conventionally used standard errors. When corrected for within-group correlation, however, the large increase in the standard errors renders these coefficients insignificant. The coefficient of the BLS industry rate is slightly larger than that of the gender-specific rate (0.014 and 0.013, respectively). In contrast to the findings for women, the coefficient on the occupational risk measure, although positive, is not significantly different from zero in any specification.

Restricting the sample further to blue-collar men paid hourly, however, reveals a positive and significant effect (at the 5-percent level in 1-sided tests) of gender-specific occupational risk on wages, but the magnitude of the effect is about one-third that of industry risk for males as well as about one-third of the coefficient of occupation risk estimated for women. Based on the heteroskedasticity-corrected standard errors, the results indicate a significantly positive wage-risk trade-off for industry risk, but once again there is a large increase in the robust standard errors corrected for within-group correlation which renders these insignificant.

V. Implicit Value of an Injury or Illness

The preceding results demonstrate that women and blue-collar men receive a significant compensating differential for job risk. Table 4 provides estimates of the wage premia for bearing risk and the implicit value of an injury or illness based on the estimates presented in Table 3. The wage premium per unit of risk is $\partial w/\partial q$, where w is the hourly wage used in the estimation of the wage equations, and q represents the risk measure used. Evaluated at the sample means of risk and hourly wages, the compensation for female workers for average risk is 1.9–2.6 percent of hourly wages. Assuming 2,000 hours worked per year, female workers at the average risk level earn a wage premium of \$408–\$563 annually.

Since the injury and illness incidence rates are per 100 full-time workers, the implicit annual value of an injury or illness is calculated as

$$(5) \qquad \partial w/\partial q \times 100 \times 2{,}000,$$

again assuming 2,000 hours worked per year for full-time employment. Based on the female-specific industry rate, the implicit value of an injury or illness is around \$20,000, while the estimated value based on the occupation risk measure yields an implicit value around \$30,000.

For comparison, Table 4 also presents corresponding values for blue-collar men. The values based on industry risk indicate that men's higher average risk level results in a larger annual compensation for risk. The values based on the occupational risk measure reveal similar values of annual compensation for female workers and for male hourly blue-collar workers. The implicit annual value of an injury or illness of about \$30,000 is close to that obtained for women based on the occupational risk measure.

VI. Conclusion

Women have largely been excluded from analyses of compensating differentials for job risk since they are predominantly employed in safer, white-collar occupations. New data reveal that their injury experience is considerable. One-third of the total injury and illness cases with days away from work accrue to female workers. Adjusted for employment, women are 71 percent as likely as men to experience an injury or illness.

As one would predict on theoretical grounds, these risks generate compensating differentials. Based on gender-specific injury incidence rates for both industry and occupation, I find strong evidence of compensating wage differentials for the job risk faced by female workers. Furthermore, all women—not only women in the riskier blue-collar jobs—receive a substantial and statistically significant premium for bearing job risk. Occupational risk has a larger impact on the wage than industry risk, and when both risk measures are included in the wage equation, only occupational risk is significant. In contrast, there is a negative relation between risk and

TABLE 4—ANNUAL VALUES OF RISK COMPENSATION[a]

	All female		Male blue-collar		Male hourly blue-collar	
Risk measure:	Female industry rate	Female occupation rate	BLS industry rate	Male industry rate	Male industry rate	Male occupation rate
Annual compensation for average risk[b]	\$408	\$563	\$1,067	\$1,091	\$1,221	\$567
Risk differential as a percentage of average wage[c]	1.93	2.66	4.92	5.03	5.77	2.68
Implicit annual value of an injury or illness[d]	\$19,631	\$29,023	\$29,485	\$27,750	\$31,148	\$10,453

[a] Based on coefficient estimates in Table 3.
[b] Calculated as $q \times \partial w/\partial q \times 2{,}000$, where q denotes the risk measure.
[c] Calculated as $(q \times \partial w/\partial q)/w$.
[d] Calculated as $\partial w/\partial q \times 100 \times 2{,}000$.

earnings for white-collar men. This is a puzzling finding, since the use of occupation-specific incidence rates reduces the measurement error that may result from imputing industry risk averages to men in safer white-collar occupations.

In contrast to the estimates based on gender-specific risk measures, estimates based on the BLS industry rate fail to reveal evidence of a compensating differential for job risk faced by women. Imputing this measure of overall industry risk to female workers apparently results in measurement error too great to yield reliable estimates of the wage-risk trade-off for female workers.

The wage-risk trade-off and the implicit value of an injury or illness are of a magnitude similar to that found in this study for male blue-collar workers. Since women comprise over 45 percent of the labor force, it is comforting to discover that, at least with regard to job risk, women and blue-collar men face a wage-determination process yielding similar compensation for job risk.

REFERENCES

Huber, Peter J. "The Behavior of Maximum Likelihood Estimates Under Non-standard Conditions," in Lucien M. Le Cam and Jerzy Neyman, eds., *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Berkeley, CA: University of California Press, 1967, pp. 221–33.

Rogers, William H. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin*, 1993, *13*, pp. 19–23.

StatCorp. *Stata statistical software: Release 5.0*. College Station, TX: Stata Corporation, 1997.

U.S. Department of Commerce, Bureau of the Census. *Census of population and housing, 1990*, [United States]: *Public use microdata sample: 5-percent sample* [computer file]. Washington, DC: U.S. Department of Commerce, Bureau of the Census [producer], 1993.

______. *Current population survey,* March 1994: Annual Demographic Survey. Washington, DC: U.S. Department of Commerce, Bureau of the Census, 1994.

U.S. Department of Labor, Bureau of Labor Statistics. *Survey of occupational injuries and illnesses in the United States, 1993*. Washington, DC: U.S. Government Printing Office, 1995a.

______. *News release USDL-95-142*. Washington, DC: U.S. Government Printing Office, 1995b.

Viscusi, W. Kip. "The Value of Risks to Life and Health." *Journal of Economic Literature*, December 1993, *31*(4), pp. 1912–46.

Did Unilateral Divorce Raise Divorce Rates? Evidence from Panel Data

By LEORA FRIEDBERG*

Since early 1996, at least eight states have begun to consider rolling back unilateral and no-fault divorce—a major reversal of the liberalizing trend in divorce laws which began around 1970.[1] The proponents of tightening the divorce regime often argue that making divorce more difficult will strengthen families. However, both theoretical and empirical evidence is decidedly mixed over whether the "no-fault revolution" actually contributed to the sharp increase in divorce rates in the United States observed over the last 30 years.

Thorough empirical work on the subject by H. Elizabeth Peters (1986) was followed by a comment by Douglas W. Allen (1992) and a reply by Peters (1992). Using the same cross section of women observed in 1979, Peters found no role for the switch to unilateral divorce in explaining rising divorce rates, while Allen found a significant impact. The differences in their results centered on controls for geographical heterogeneity in divorce propensities. In this paper I address the dispute by using a panel of state-level divorce rates. The sample includes virtually every divorce in the United States over the entire duration of the law changes and allows thorough controls for heterogeneity in divorce propensities across states and over time—which turn out to be crucial to the results. Including state-specific trends allows unobserved state divorce propensities to trend linearly and even quadratically over time and reveals that unilateral divorce raised divorce rates significantly and strongly.

An additional difficulty in this literature arises in defining precisely what is unilateral divorce. Peters and Allen disagreed whether separation requirements qualify as unilateral divorce. Margaret F. Brinig and F. H. Buckley (1998) questioned whether divorce in states that retain fault grounds for property settlement can be considered unilateral. The empirical analysis in each of those papers retained a binary classification of divorce laws, yet the impact of different types of unilateral divorce can be tested. I find here that the type of unilateral divorce matters. States that adopted divorce laws which were more strongly unilateral had greater increases in the divorce rate. Nevertheless, the evidence shows that adopting any type of unilateral divorce raised the divorce rate.

I combine both sets of results to arrive at final estimates of the impact of unilateral divorce. The estimates suggest that the divorce rate would have been about 6 percent lower in 1988 if no type of unilateral divorce had been adopted in those states that switched to unilateral divorce after 1968. The move toward unilateral divorce accounted for 17 percent of the increase in divorce rates between 1968 and 1988. Lastly, I demonstrate that the law changes do not appear to be endogenous, and that the effect of unilateral divorce on divorce behavior was permanent, not temporary.

I. Measuring the Impact of Unilateral Divorce

The divorce rate in the United States soared in the 1960's and 1970's, rising from 2.2 thousand people in 1960 to 5.0 in 1985, demonstrated in Figure 1 by the upper

---

* Department of Economics 0508, University of California-San Diego, 9500 Gilman Drive, La Jolla, CA 92093, and the National Bureau of Economic Research. I wish to thank Jonathan Gruber, Shoshana Grossbard-Shechtman, Jerry Hausman, James Poterba, and especially Aaron Yelowitz for many useful discussions; anonymous referees for their comments; and Margaret Brinig for sharing her data on state divorce laws.

[1] See for example John Leland (1996) in *Newsweek*.

[2] Andrew J. Cherlin (1992) documented that the divorce rate rose steadily in the late nineteenth and twentieth centuries. After a sharp peak during World II, it dropped below trend in the 1950's. The rise in the 1960's and 1970's looks especially stark compared to the low level in the 1950's, but it also was steeper than the century-long trend.

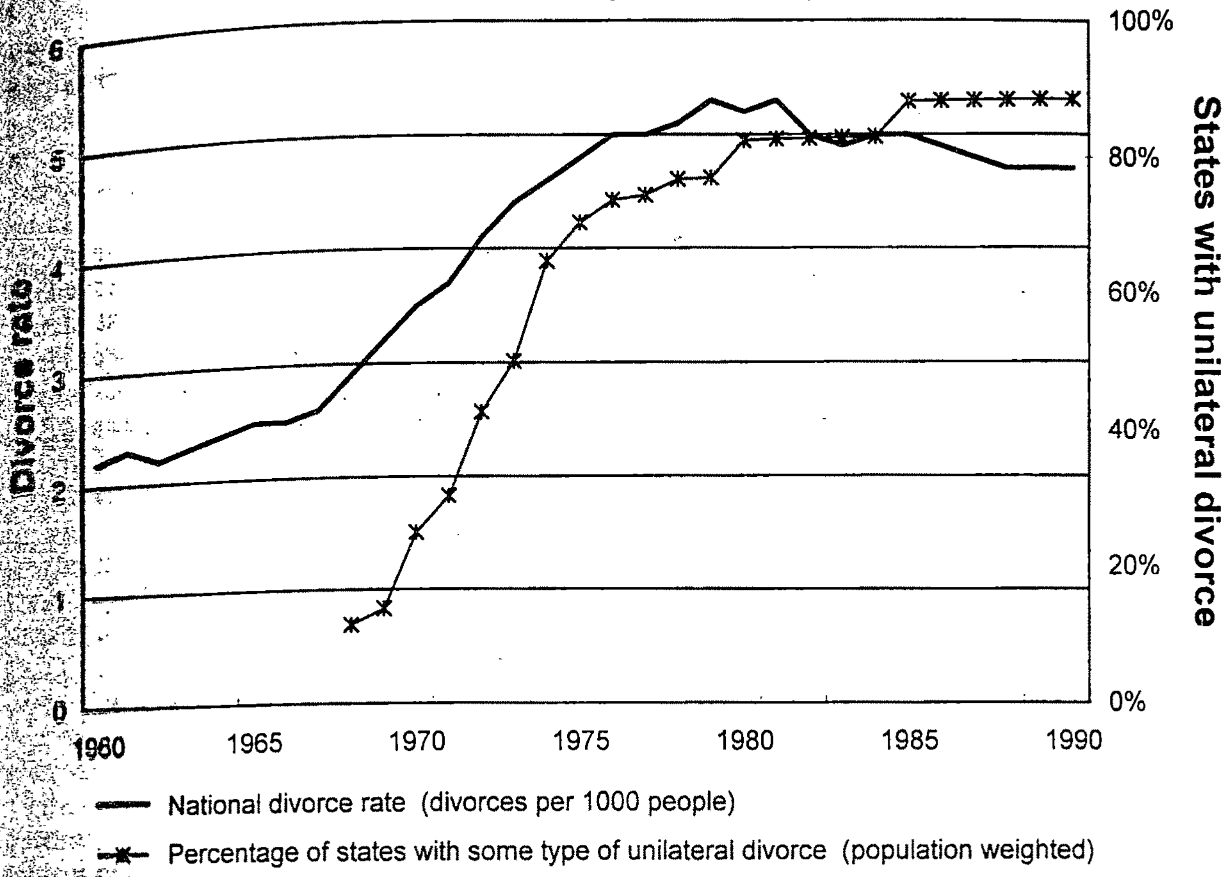


FIGURE 1. THE DIVORCE RATE AND DIVORCE LAWS

During that time, states substantially liberalized and simplified their divorce laws. For example, in the absence of fault most states used to require mutual divorce, to which both spouses had to consent. Many states abandoned mutual-divorce requirements in favor of unilateral divorce, which required the willingness of only one spouse to divorce. Most states also adopted some form of no-fault divorce, eliminating the need for one spouse to prove a transgression by the other. The lower line in Figure 1 shows the rise in the population-weighted percentage of states with unilateral divorce starting in 1968.[3] Casual observation suggests attributing at least some of the increasing prevalence of divorce to the widespread shift to unilateral and no-fault divorce, which appears to make divorce simpler. On the other hand, lawmakers and legal experts at the time argued that the motivation for changing the laws was to bring law books into alignment with actual practice and to eliminate the need to assign blame during a divorce, but not to make divorce easier.[4]

A theoretical model of divorce laid out by Gary S. Becker (1981) and Peters (1986) showed that, under perfect information and costless bargaining, a marriage will be dissolved when it is efficient to do so: ". . . when the joint value of the marriage is less than the sum of the values of opportunities that face each spouse at divorce'' (Peters, 1986 p. 438). Barring asymmetric information and transactions costs, a change in the law from mutual to unilateral divorce would not make divorce more likely. By the Coase theorem, it would simply redistribute property rights from the spouse who wished to remain in the marriage to the spouse who wished to leave. Thus, under mutual divorce the spouse who wished to leave the marriage had to compensate the spouse who wished to stay in order to obtain the efficient outcome of divorce, while under unilateral divorce the spouse who wishes to stay in the marriage must compensate the one who wishes to leave. The law change would alter the property rights and resulting compensation scheme between the spouses without making them more likely to divorce.

---

[3] Figure 1 shows the percentage of states with any type of unilateral divorce broadly or narrowly defined. Section [illegible] discusses the classification of unilateral-divorce laws.

[4] This intention is stated in the letter of the law in several states, such as Wisconsin. With respect to actual practice at the time, studies showed that many divorces did not involve real transgressions, regardless of legal requirements (Doris Jonas Freed and Henry H. Foster, Jr., 1979). As an example, news stories several years ago reported that, while the ex-wife of former Governor Doug Wilder of Virginia had divorced him on the grounds of cruelty, they had fabricated the grounds in order to obtain a divorce which they both wanted. Similar anecdotes suggested to the legal and legislative community that blame was not necessary to determine whether a couple should divorce.

A. *Estimation Approach*

The typical test of whether the Coase theorem applies, and hence the type of divorce law does not matter, asks whether people living under unilateral-divorce laws are no more likely to divorce than people living under mutual-divorce laws. Peters (1986) and Allen (1992) took this approach, using the same cross-sectional data but arriving at different conclusions. Both used a supplement from the 1979 Current Population Survey of individuals' marital histories. They estimated a logit on a sample of those women who reported being married in 1975, where the left-hand-side variable was one if a woman was divorced in 1978 and zero if she stayed married. Controlling for demographic variables believed to affect the likelihood of divorce, they used the variation of divorce regimes across states in 1978 to identify whether unilateral divorce raised divorce rates.

Peters and Allen disagreed about including controls for geographic differences in divorce propensities. Peters estimated a coefficient on unilateral divorce of virtually zero. Allen demonstrated that, when region dummies and a state's 1970 divorce rate were excluded from the regression, the coefficient on the law became significant and of an important magnitude, implying that unilateral divorce raises the probability of divorce by 1.4 percent. He suggested that the region dummies did not serve a purpose in explaining divorce independe but simply absorbed much of the variatio divorce laws.[5]

With panel data, the role of cross-st heterogeneity in divorce behavior can be e plored in more detail, without having to b specific about the sources of heterogeneit State fixed effects control for unobserved in fluences on divorce that vary across states so that the effect of the divorce law is ide tified from its variation within a state ov time. For this purpose, I assembled a pane of state-level divorce rates over a 20-yea period. This data records virtually every di vorce in the United States.[6]

The relationship which Peters estimated ca be easily adapted for a panel of state divor rates. Her regression took the following for

$$(1)\qquad \text{divorce}_{is} = a_0 + a_1 * \text{unilateral}_s + \mathbf{Z}_{is} * \mathbf{a}_2 + e_{is},$$

which describes a population of women wh are initially married. Each observation is su scripted for the individual *i* and the state i which she lives. The variable *divorce* is one i the woman is divorced when observed an zero if she is not. Of interest is the varia *unilateral,* which is one if the woman's st has a unilateral-divorce law and zero if it h a mutual-divorce law. Demographic chara teristics which may affect the propensity to d vorce, such as education, number of childr and age, are included in **Z**.[7]

---

[5] Allen also disagreed that the Coase theorem cou ply to this situation, arguing that marriage is not a with perfectly defined property rights and that divorc volves transaction costs. This disagreement does n terfere with the empirical approach of measuring impact of unilateral-divorce laws on divorce propens

[6] Using a longitudinal survey like the Panel St Income Dynamics might be preferable, since it has individual-level information on the correlates of di However, it has the disadvantages of a very small s size and potential bias due to sample attrition bein related with divorce.

[7] Peters estimated a logit, not a linear probability as expressed in (1). She used age, education, race, of kids under 18 added linearly and squared, a dum living in a Standard Metropolitan Statistical Area, dummy for being in the labor force as right-ha variables.

To estimate the relationship with state-level time-series data, add up equation (1) over observations within a state and divide by population to obtain the following:

$$(2)\quad \text{divrate}_{st} = b_0 + b_1 * \text{unilateral}_{st} + b_{2s} * \text{state}_s + b_{3t} * \text{year}_t + u_{st}.$$

Now, the dependent variable *divrate* is the state-level divorce rate, which is the number of divorces that occur within a state each year divided by the state population in thousands.[8] The disturbance u is equal to $(\Sigma_{i \in s} e_{ist})/\text{pop}_{st}$, the sum of all individual disturbances within a state each year divided by the population. This makes u heteroskedastic, so population is used to perform weighted least squares.[9] Using a panel allows the inclusion of *year*, which are year effects that control for evolving unobserved national attributes that affect the likelihood of divorce.

The demographic characteristics **Z** from (1) become, for example, the number of urban residents or the number of children in the state. Equation (2) assumes that the demographic characteristics are unchanging within a state over time. The coefficients cannot be separately identified, so the state characteristics are subsumed in the fixed-effect *state*. The use of state effects, which is only possible with a panel, is advantageous because it explains divorce patterns flexibly, rather than having to include all relevant divorce covariates explicitly. However, the factors which influence divorce may vary within a state over time, confounding the estimates of the state effects. That will also bias the estimate of the coefficient on *unilateral* if the changing factors are correlated with the law changes across states and if such factors do not change at a national level uniformly and get picked up by the year effects. The specification in (2) can be relaxed further to allow for such changing influences within a state over time, as follows:

$$(3)\quad \text{divrate}_{st} = c_0 + c_1 * \text{unilateral}_{st} + c_{2s} * \text{state}_s + c_{3t} * \text{year}_t + c_{4s} * \text{state}_s * \text{time}_t + c_{5s} * \text{state}_s * \text{time}_t^2 + u_{st}.$$

Time is a time trend, so the interaction terms *state * time* and *state * time*^{2} are linear and quadratic trends for each state. They capture quadratic trends in state-level characteristics that influence divorce, with the slopes of the trends allowed to vary across states.[10] Thus, (3) represents an extremely flexible way to control for heterogeneous divorce behavior, identifying the impact of *unilateral* from a jump in the pattern of divorce which is distinguishable from a smooth quadratic. While it remains possible that this specification still does not capture all of the factors explaining *divrate*, additional evidence against the endogeneity of *unilateral* will be discussed later.

B. *Data*

The longitudinal data on state divorce rates cover all 50 states and the District of Columbia from 1968 to 1988. The "no-fault revolution" began in 1970 when California replaced all fault grounds for divorce—such as adultery and cruelty—with the sole ground of irretrievable breakdown of the marriage. Over the following decade and a half, every state legalized no-fault grounds for divorce. As reported in Table 1, most states adopted unilateral divorce as well, under which the no-fault grounds do not require the agreement of both spouses.[11]

---

[8] To be precise, *divrate* should measure the number of divorces per thousand married women, but that information is not readily available. To get at the true divorce rate, I used the sampling weights to aggregate data from the *Current Population Survey* (U.S. Department of Commerce, Bureau of the Census) and obtain estimates of the stock of married women by state and year, in order to normalize the measured divorce rate. This had the effect of strengthening the estimated impact of *unilateral* on divorce rates, but not significantly. In the rest of the paper I use the available measure of the divorce rate, principally because the *Current Population Survey* did not report explicit state identifiers for numerous states before 1978.

[9] The intuition behind weighted least squares is that a negative effect of *unilateral* in, say, California will carry more weight than a positive effect in New Hampshire.

[10] The use of state-specific linear trends follows Louis S. Jacobson et al. (1993).

[11] The information in Table 1 is derived from Freed and Foster (1977, 1979, 1981), Harvey J. Sepler (1981), Freed and Timothy B. Walker (1990), and Brinig and Buckley (1998).

TABLE 1—DIVORCE LAWS, BY STATES

State	(1) Unilateral divorce	(2) Unilateral divorce, including separation requirements	(3) Unilateral divorce, no-fault settlement
Alabama	✓	✓	no, fault is considered in the settlement (f.s.)
Alaska	✓	✓	✓
Arizona	✓	✓	✓
Arkansas	no	no, separation grounds allowed but separation must be mutually agreed	no
California	✓	✓	✓
Colorado	✓	✓	✓
Connecticut	✓	✓	no, f.s.
Delaware	no	no	no
District of Columbia	no	✓, 1-year separation requirement (s.r.)	no, f.s.
Florida	✓	✓	no, f.s.
Georgia	✓	✓	no, f.s.
Hawaii	✓	✓	✓
Idaho	✓	✓	no, f.s.
Illinois	no	✓, 2-year s.r.	✓
Indiana	✓	✓	✓
Iowa	✓	✓	✓
Kansas	✓	✓	no, f.s.
Kentucky	✓	✓	✓, 1987
Louisiana	no	✓, 1-year s.r.	no, f.s.
Maine	✓	✓	✓
Maryland	no	✓, 5-year s.r.; later 2	no, f.s.
Massachusetts	✓	✓	no, f.s.
Michigan	✓	✓	no, f.s.
Minnesota	✓	✓	✓
Mississippi	no	no	no
Missouri	no	✓, 2-year s.r.	✓
Montana	✓	✓	✓
Nebraska	✓	✓	✓
Nevada	✓	✓	✓
New Hampshire	✓	✓	no, f.s.
New Jersey	no	✓, 18-month s.r.	no, f.s.
New Mexico	✓	✓	✓
New York	no	no	no
North Carolina	no	✓, 1-year s.r., need agreement on alimony	no, f.s.
North Dakota	✓	✓	no, f.s.
Ohio	no	✓, 1-year s.r.	no, f.s.

TABLE 1—*Continued.*

State	(1) Unilateral divorce	(2) Unilateral divorce, including separation requirements	(3) Unilateral divorce, no-fault settlement	(4) Since
Oklahoma	✓	✓	✓, 1975	pre-1968
Oregon	✓	✓	✓	1973
Pennsylvania	no	✓, 3-year s.r., and court says marriage broken	no, f.s.	1980^a
Rhode Island	✓	✓	no, f.s.	1976
South Carolina	no	✓, 3-year s.r.; later 1	no, f.s.	1969
South Dakota	✓	✓	no, f.s.	1985
Tennessee	no	no, 2-year s.r. if no kids	no	—
Texas	✓	✓	no, f.s.	1974
Utah	no	✓; 3-year s.r.	no, f.s.	pre-1968
Vermont	no	✓, 6-month s.r.	no, f.s.	pre-1968
Virginia	no	✓, 2-year s.r.	no, f.s.	pre-1968
Washington	✓	✓	✓	1973
West Virginia	no	✓, 2-year s.r.; later 1	no, f.s.	pre-1968
Wisconsin	no	✓, 1-year voluntary s.r.; if one party disputes it, court decides if differences are irreconcilable	✓	1977^a
Wyoming	✓	✓	no, f.s.	1977

Notes: (1) ✓, state has unilateral divorce. Unilateral divorce requires the consent of only one spouse and is granted on no-fault grounds. No-fault grounds include irretrievable breakdown, irreconcilable differences, and/or incompatibility.

(2) ✓, state has unilateral divorce, and the no-fault grounds now include or require separation for a required length of time.

(3) ✓, state has unilateral divorce, and also does not allow fault grounds to be considered for property settlement and alimony.

(4) Year when unilateral divorce instituted. If the law change occurred in the second half of the year, then it is attributed to the following year.

Sources: Sepler (1981), Peters (1986), Freed and Walker (1990), Brinig and Buckley (1998).

[a] Definition of unilateral divorce is disputable. The estimation results were not found to be sensitive to the classification of these states.

Table 1 also distinguishes among alternative categorizations of divorce laws. For comparability, the results in this section use Peters' classification of unilateral divorce, which excludes those states where the only unilateral ground is living separate for a length of time. Later I relax the restriction that separation requirements have no effect on divorce rates.

Divorce rates by state, reported in *Vital Statistics of the United States* (National Center for Health Statistics, 1960–1988), are shown in Figure 2. The sample covers virtually every divorce in the country for 20 years. It is not complete because, in certain states in certain years, *Vital Statistics* reports that the data was collected in an irregular or incomplete fashion. A set of dummies, listed in the Appendix, control for those instances. A few missing values are also listed in the footnotes to Table 2.

C. *Results*

The coefficient on *unilateral* turns out to be quite sensitive to the nature of state and time controls used, as Allen demonstrated. The results of estimating the basic regression and then adding state and year effects, as in equation (2), are shown in Table 3.

TABLE 2—DIVORCE RATES, BY STATES

State	Average divorce rate: Divorces per 1,000 people		
	1968–72	1973–77	1978–82
Alabama	4.4	6.1	6.7
Alaska	5.6	7.6	8.5
Arizona	7.0	7.4	7.3
Arkansas	5.5	6.9	6.9
California	4.9	5.9	5.6
Colorado	4.7	6.4	6.3
Connecticut	2.1	3.6	4.3
Delaware	2.9	4.6	5.0
District of Columbia	3.3	4.6	6.2
Florida	5.8	7.2	7.2
Georgia	4.0	5.7	6.2
Hawaii	3.6	4.9	4.7
Idaho	4.9	6.2	6.8
Illinois	3.4	4.3	4.5
Indiana[a]	4.6	6.1	7.0
Iowa	2.6	3.5	3.9
Kansas	3.8	5.2	5.5
Kentucky	3.3	4.2	4.3
Louisiana[b]	2.7	3.2	3.8
Maine	3.7	4.9	5.5
Maryland	2.4	3.7	4.0
Massachusetts	2.1	2.8	3.1
Michigan	3.4	4.6	4.7
Minnesota	2.2	3.3	3.8
Mississippi	3.6	5.0	5.4
Missouri	3.9	5.0	5.5
Montana	4.4	5.7	6.2
Nebraska	2.5	3.6	4.1
Nevada	20.9	17.2	16.2

TABLE 2—*Continued.*

	Average divorce rate: Divorces per 1000 people			
State	1968–72	1973–77	1978–82	1983–88
New Hampshire	3.5	5.0	5.5	4.8
New Jersey	1.7	2.8	3.6	3.6
New Mexico[c]	4.3	7.3	7.8	7.8
New York	1.5	3.0	3.5	3.7
North Carolina	2.8	4.0	4.8	4.9
North Dakota	1.7	2.7	3.3	3.4
Ohio	3.6	4.9	5.3	4.8
Oklahoma	6.3	7.6	7.8	7.4
Oregon	4.7	6.3	6.7	5.8
Pennsylvania	2.0	2.9	3.3	3.3
Rhode Island	1.6	3.0	3.7	3.8
South Carolina	2.3	3.3	4.3	4.1
South Dakota	2.1	3.2	3.9	3.7
Tennessee	4.2	6.1	6.5	6.3
Texas	4.7	5.9	6.6	5.9
Utah	3.8	4.9	5.3	5.1
Vermont	2.2	3.8	4.6	4.5
Virginia	2.6	3.8	4.5	4.3
Washington	5.3	6.9	6.8	6.0
West Virginia	3.0	4.5	5.3	5.0
Wisconsin	2.0	2.9	3.7	3.5
Wyoming	5.6	7.1	8.0	7.2

Note: Averages across years are computed using annual state population for weights.
Source: Vital Statistics of the United States (National Center for Health Statistics, 1960–1988).
[a] Data are for 1968–69, 1974, 1976–87.
[b] Data are for 1971–72, 1976, 1978–80, 1982–83.
[c] Data are for 1970, 1974–80, 1983–85.

Regression 3.1 has no state or year effects. The estimated coefficient on *unilateral* is statistically significant and very large, at 1.802 ([illegible]), with an adjusted R^2 of 0.314. Similar results are obtained with year effects in Regression 3.2. Regression 3.3 adds state fixed effects, which are meant to absorb everything affecting divorce behavior at the state level that is constant over time. Adding state effects drives down the coefficient on *unilateral* to 0.004 (0.056), while raising the explanatory power of the regression to 0.946.

TABLE 3—REGRESSION RESULTS

	Dependent variable: Divorce rate (divorces per 1,000 people)				
Independent variables:	3.1	3.2	3.3	3.4[a]	3.5
Unilateral	1.802 (0.087)	1.509 (0.090)	0.004 (0.056)	0.447 (0.050)	0.441 (0.0
Adjusted R^2	0.314	0.362	0.946	0.976	0.982
Year effects[b]	No	Yes, $F = 4.9$	Yes, $F = 89.0$	Yes, $F = 95.3$	Yes, $F = 8$
State effects[b]	No	No	Yes, $F = 217.3$	Yes, $F = 196.2$	Yes, $F = 1$
State trend, linear[b]	No	No	No	Yes, $F = 24.7$ State trends and effects, $F = 256.0$	Yes, $F = 9$
State trend, quadratic[b]	No	No	No	No	Yes, $F = 6$ State trends and effe $F = 224$

Notes: Regressions of state divorce rates, 1968–88, on whether a state has a unilateral-divorce law. Unilateral-div law is defined according to column (1) of Table 1. Standard errors are in parentheses. Regressions are weighted by population. $N = 1{,}043$. Data for some states in certain years is missing. For others it is incomplete, which is acco for by a set of dummies described in the Appendix.

[a] The coefficients on the state effects and trends are shown in Tables 6 and 7.

[b] All reported F-statistics have corresponding p-values that are smaller than 0.00005.

The state effects appear to explain most of the pattern of divorce, which had previously been picked up by the divorce law.

Adding State-Specific Trends.—The results above make it clear that unobserved covariates and unobservable divorce propensities—which may include, for instance, social attitudes, religious beliefs, and family size—are the main determinants of divorce. However, assuming that they are constant within a state over 20 years or changing over time but uniform across the nation is restrictive. State-year interactions are completely unrestrictive but not feasible. State-specific trends, though, are a feasible alternative that allow the unobserved state factors influencing divorce to have a linear trend, with the trend allowed to vary across states. The estimated coefficient on *unilateral* retains the same interpretation as before: a positive coefficient measures by how much a unilateral-divorce law raises a state's divorce rate, which now might also have a positive or negative trend.

Regression 3.4 includes the state-specific trends. The state trends are jointly significant and raise the adjusted R^2 to 0.976. More portantly, the coefficient on *unilateral* is larger and has a smaller standard error. Th estimate of 0.447 (0.050) means that uni eral divorce raised the divorce rate by than 0.4 divorces per thousand people, a stantial impact relative to the average div rate over the period of 4.6.

Why State-Specific Trends Matter.—Why big change in the estimate of *unilateral's* efficient when state trends are added? It is cause imposing a flat state divorce prop misspecifies the underlying divorce beha and in turn camouflages the variation ind by the law change. Naturally, imposing stant state divorce propensities when they really trending will bias the estimates of constants—the estimated intercepts will flect the average of the trend instead of the intercept, which is the state's initial div propensity only. If, further, the omitted s are correlated with the law changes, the is omitted variable bias. The resulting bias be depicted visually, as in Figures 2A and The upper line in Figure 2A is California

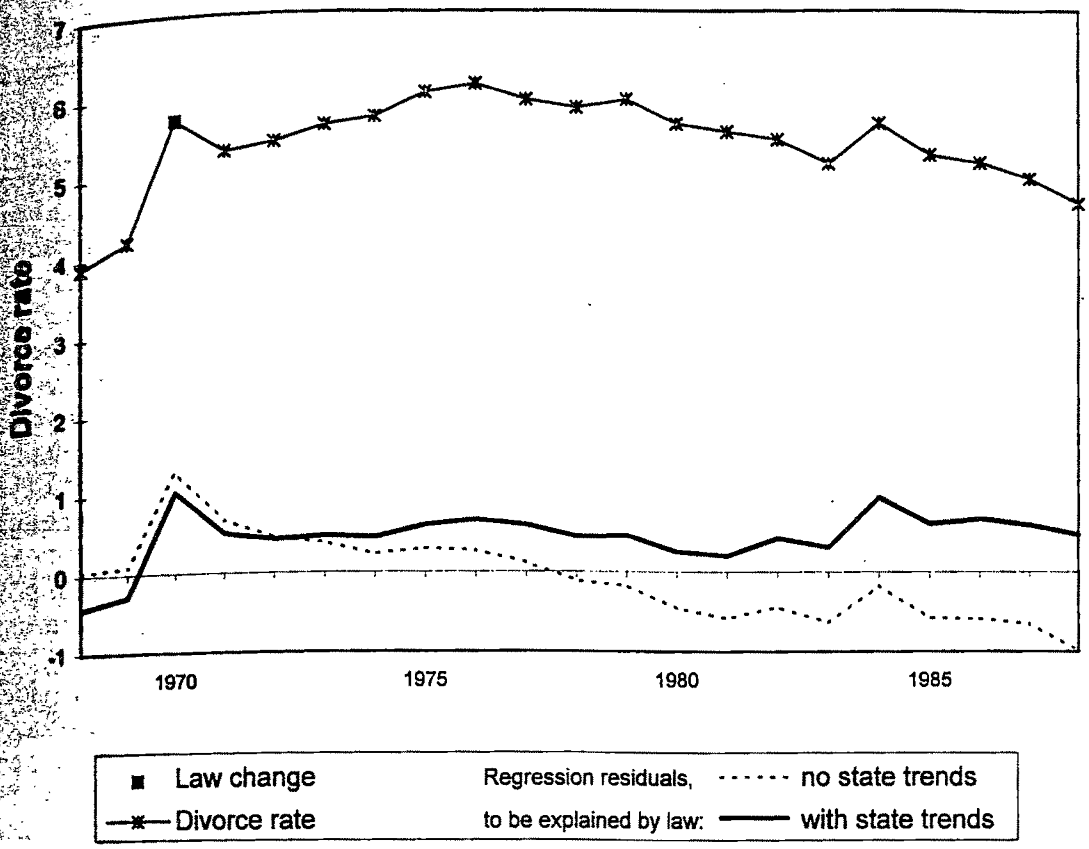


FIGURE 2A. RESIDUALS FROM REGRESSIONS EXPLAINING THE DIVORCE RATE IN CALIFORNIA

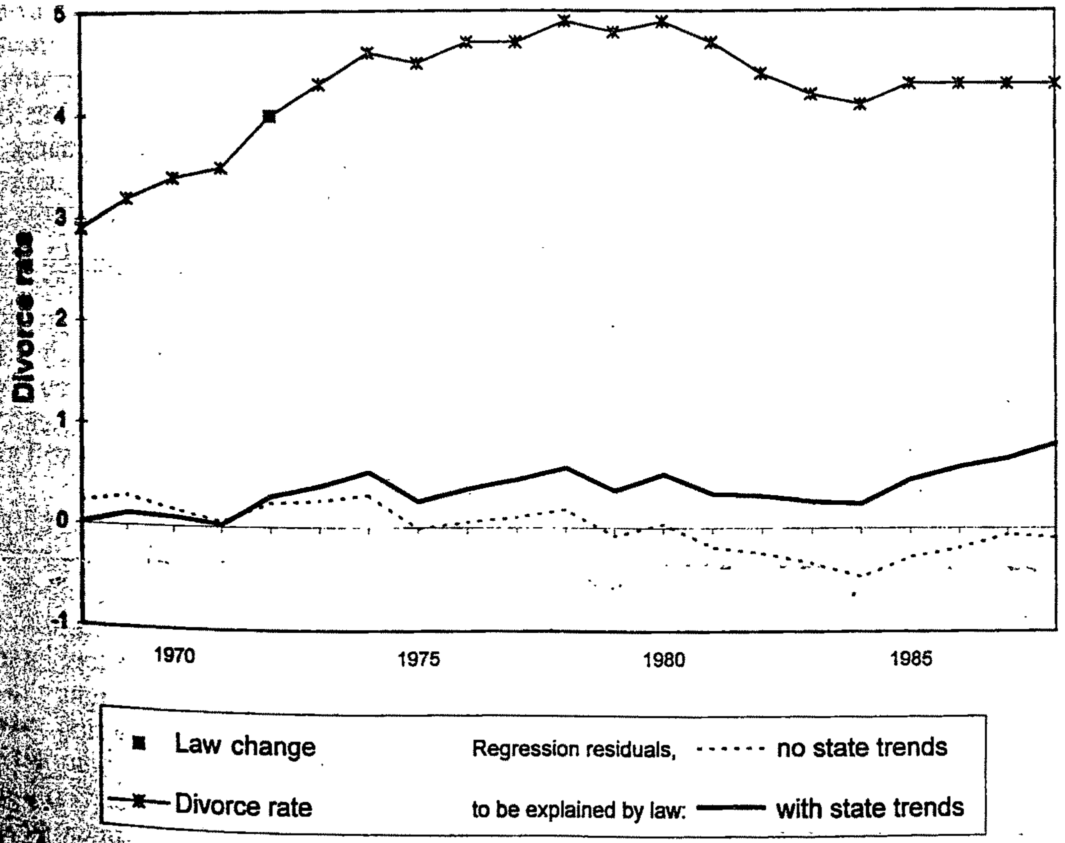


FIGURE 2B. RESIDUALS FROM REGRESSIONS EXPLAINING THE DIVORCE RATE IN MICHIGAN

tual divorce rate, with the box signaling the year of the law change. The lines near the bottom show what remain of the divorce rate to be explained by the law change in each regression. The dashed line removes the constant and the state and year effects from regression 3.3, and the solid line removes the (negative) state trend as well from regression 3.4. Both lines jump in the year of the law change, but the dashed line from regression 3.3 ultimately confounds the influence of *unilateral* with the omitted downward trend in California's divorce propensity, biasing downward the estimated effect of *unilateral.* The solid line, on the other hand, demonstrates that, controlling for all other influences, the law change raised the divorce rate permanently. Figure 2B illustrates the same point for Michigan. While the dashed line obscures the effect of *unilateral,* the solid line shows that the law change raised the divorce rate.

The possibility remains that the unobservables exhibit more complex dynamic behavior than can be captured by the linear trend. Therefore, regression 3.5 adds a quadratic in the time trend for each state, which places stronger restrictions on the kind of variation in the divorce rate that can be attributed to the law change. Now, the effect of *unilateral* is identified from a break in the pattern of divorce that is distinguishable from a smooth quadratic. The state quadratic terms in regression 3.5 are jointly significant and raise the adjusted R^2. Meanwhile, the estimated coefficient on *unilateral* is virtually unchanged at 0.441 (0.055), compared to 0.447 (0.050) without the quadratics. The estimate of *unilateral*'s impact is robust even when the underlying divorce behavior is parameterized very flexibly.

The strongly significant coefficient on *unilateral* when state trends are included indicates that both Peters' and Allen's regressions were misspecified. Unless the individual-level demographic controls he used fully explain the divorce rate, Allen—by omitting any controls for state divorce propensities—was naturally more likely to find an impact of the law change, as regression 3.1 demonstrated. Yet, Peters' approach of including the state's 1970 divorce rate was similar to including only the state fixed effect as in regression 3.3, obscuring the impact of *unilateral.*

II. What Is Unilateral Divorce?

A. *Classification of State Divorce Laws*

The theory regarding unilateral divorce is concerned simply with whether one party can end the marriage without the consent of the other, but what constitutes unilateral divorce in practice is not so clear-cut. One reason that classifying state laws requires some care is that most of the legal scholarship has focused on whether divorce grounds are fault or no-fault, rather than mutual or unilateral. However, no-fault divorce is not synonymous with unilateral divorce, which is the subject of the Coase theorem. A few states (such as Delaware and Arkansas) allow no-fault grounds but require mutual agreement on them and are classified as having mutual divorce.

Two further issues arise in the classification of state laws, both of which I will test for. Some states allow unilateral divorce only if the couple has lived separately for a length of time. Peters classified those states as having mutual divorce. She argued that separation requirements impose a large cost to divorce unilaterally, so that it would continue to be in the interest of the spouse who wishes to leave the marriage to compensate the spouse who wishes to remain in the marriage in order to divorce promptly. Allen reclassified separation states as having unilateral divorce but found little change in the estimated effect of *unilateral.* This issue is addressed here more thoroughly by allowing for the two types of unilateral divorce to have different effects on the divorce rate. According to Peters' hypothesis, separation requirements will have an insignificant effect on the divorce rate, compared to other no-fault grounds. The strategy employed here allows separation requirements to have a smaller, but still significant, effect on divorce.

A final issue raised by Brinig and Buckley concerns the grounds for property settlements and alimony. Some states grant no-fault divorces but still allow fault to be considered in the division of property and determination of alimony. Brinig and Buckley argue that fault considerations in property division raise the cost of divorce for the spouse who wishes to leave, in effect keeping the property rights with the spouse who wishes to stay in the marriage. Brinig and Buckley classify states as no-

TABLE 4—ALTERNATIVE CLASSIFICATIONS OF DIVORCE LAWS

		Dependent variable: Divorce rate (divorces per 1,000 people)			
		3.4	4.1	4.2	4.3
Unilateral		0.447 (0.050)	0.450 (0.050)	0.545 (0.068)	0.549 (0.067)
Unilateral defined as:					
Grounds for divorce: Other no-fault, not including separation		✓	✓	✓	✓
Grounds for property settlement	No-fault	✓	✓	✓	✓
	Fault	✓	✓	—	—
Grounds for divorce: Separation		—	0.167 (0.059) [0.0002]	—	—
Grounds for property: Fault		—	—	0.392 (0.056) [0.033]	—
Grounds for divorce: Other no-fault, not including separation Grounds for property: Fault		—	—	—	0.396 (0.056)
Grounds for divorce: Separation Grounds for property: No-fault		—	—	—	0.133 (0.091)
Grounds for divorce: Separation Grounds for property: Fault		—	—	—	0.192 (0.078)
Adjusted R^2		0.976	0.976	0.976	0.976

Notes: See the notes to Table 3 for more detail. See the notes to Table 1 for a discussion of the classification of state divorce laws. Standard errors in parentheses. Reported in brackets are p-values of the F-statistics testing the equality of the two coefficients.

only if they have no-fault property division as well.[12] I test their hypothesis explicitly by allowing separate effects on divorce for states that have unilateral divorce with no-fault property division or with fault property division.

B. *Estimating the Impact of Different Unilateral-Divorce Laws*

In Table 4, I report the results of including separate variables for the different types of unilateral divorce. Regression 4.1 explores the impact of unilateral divorce with separation requirements. If separation requirements raise the cost of divorce, in effect keeping some of the property rights with the spouse wishing to stay in the marriage, it would be reflected in a smaller or insignificant coefficient. In regression 4.1, the impact of both types of unilateral divorce is statistically significant. The coefficient on unilateral divorce with separation requirements of 0.167 (0.059) is statistically different from the coefficient of 0.450 (0.050) on strict unilateral divorce, as indicated by the reported p-value of the F-test. Therefore, the effect of unilateral divorce with separation requirements is substantially smaller but still positive.[13]

---

[12] Another condition for no-fault divorce in Brinig and Buckley's classification is that a state not retain any fault grounds for divorce. This does not seem to be relevant to the existence of unilateral divorce, however.

[13] I estimated a regression similar to 4.1 but allowing distinct effects of different required lengths of separation. However, there is not enough variation among states in the lengths required to identify them separately. The coefficient on a 12-month separation requirement is

Regressions 4.2 and 4.3 allow a distinct effect of unilateral divorce depending on whether grounds for property division are no-fault or fault. Even if divorce grounds are no-fault, fault considerations in property settlement and alimony may strengthen the position of the spouse who wishes to stay in the marriage. In 4.2 the coefficient on unilateral divorce with fault property division falls to 0.392 (0.056), significantly different from the coefficient on strict unilateral divorce at a 96.7-percent confidence level. The coefficient on strict unilateral divorce, with no-fault grounds for property division, is 0.545 (0.068), somewhat larger than the coefficient of 0.447 (0.050) when grounds for property division are ignored.

Regression 4.3 combines the approaches taken in 4.1 and 4.2 by allowing for a multiple classification of unilateral divorce. In 4.3 the coefficient on very strict unilateral divorce—unilateral divorce without separation requirements and without fault grounds for property division—is reported in the first row and rises to 0.549 (0.067). It implies a powerful effect on divorce rates, while the other three types of unilateral divorce have smaller effects. The next largest coefficient is 0.396 (0.056) for unilateral divorce without separation requirements but with fault property division. Unilateral divorce with separation requirements has a smaller impact on divorce, with respective coefficients of 0.133 (0.091) and 0.192 (0.078) without fault grounds and with fault grounds for property settlement. Though not significantly different, the two coefficients are in reverse order from what would be expected.

To sum up, this section finds that the type of unilateral divorce a state adopted mattered. The strictest unilateral divorce, without separation requirements or fault considerations in property division, raised the divorce rate by 0.549 per thousand people—11.9 percent of the average of 4.6 during the sample period. As Peters and Brinig and Buckley hypothesized, weaker versions of unilateral divorce did have smaller effects on the divorce rate—but their effects were still positive. Separation requirements proved more of a constraint on divorce behavior than fault-based property division did, according to the evidence. Overall, the results strengthen conclusions about the impact of switching to any type of unilateral regime. The estimates from regression 4.3 demonstrate that the divorce rate would have been 6.4 percent lower in 1988 if no type of unilateral divorce had been adopted in those states that switched.[14] The move towards unilateral divorce after 1968 accounted for 17.1 percent of the increase in divorce rates between 1968 and 1988.

III. Additional Results

A. *Endogenous Legislation*

Any analysis of how legal changes affect behavior raises the question of endogeneity. In the case of divorce, it might be that rising divorce rates caused the law to be changed or that increasing permissiveness towards divorce, which in turn raised divorce rates, did—not that the law changes caused divorce rates to rise. Several authors argue against this.[15] Also, Peters asserts that interest groups of divorcing individuals were not part of the debate over no-fault divorce laws and that proponents of the change were primarily legal scholars. Unfortunately, it is difficult to come by instruments in order to test properly for endogeneity.[16]

An indication of whether legislation might be endogenous can be found in the correlation across states between initial divorce rates and the likelihood and chronology of the law

---

statistically significant and about the same size as the coefficient on strict unilateral divorce. The coefficients on the other lengths of separation are far from being significant. The results indicate that 12 months is not too long for the spouse who wishes to leave the marriage to wait, when the alternative is mutual divorce, but that waiting longer apparently would impose an excessive cost.

[14] Estimating the relationship with the log of the divorce rate on the left-hand side yields very similar conclusions.

[15] For example, legal scholars Freed and Foster (19[illegible] and Sepler (1981).

[16] An instrument must be correlated with the [illegible] changes but not with the divorce rate independently [illegible] must also, for the purposes here, vary across state and [illegible] time. In a time-series study, Sepler used as a control [illegible] number of articles about no-fault divorce in legal journals, which does not vary across states. Other potential instruments relate to the composition of the legislature (number of women, number of divorced people, etc.) but [illegible] ficult to obtain.

TABLE 5—THE TIMING OF THE RESPONSE TO UNILATERAL DIVORCE

	Dependent variable: Divorce rate (divorces per 1,000 people)					
	3.4	5.1	5.2	5.3	5.4	5.5
Unilateral dummy	0.447 (0.050)	0.449 (0.057)	0.468 (0.054)	0.489 (0.060)	0.536 (0.065)	0.438 (0.055)
Dummy for law change occurring:						Sample restricted to: 1968–1979
1 period later	—	0.004 (0.067)	—	—	—	—
In this period	—	—	−0.060 (0.062)	−0.076 (0.065)	−0.117 (0.069)	—
1 period ago	—	—	—	−0.055 (0.063)	−0.090 (0.066)	—
2 periods ago	—	—	—	—	−0.108 (0.062)	—
Adjusted R^2	0.976	0.976	0.976	0.976	0.976	0.983

Notes: See the notes to Table 3. Standard errors are in parentheses. Unilateral divorce is defined according to column (1) of Table 1.

changes. It does turn out that a state's 1968 divorce rate is a reasonable predictor of *whether* a state adopted strict unilateral divorce—with a correlation coefficient as high as 0.31 (0.03), depending on the type of unilateral divorce considered.[17] This result is consistent with similar findings reported in Peters (1992). However, the initial divorce rate is not correlated with *when* a state adopted unilateral divorce.[18]

A similar conclusion results from a strategy employed by Jonathan Gruber and Maria Hanratty (1995). In their regression they added a lead dummy variable for whether a law change would occur in the subsequent year. Under the null hypothesis in which causality goes from the law change to the left-hand-side variable, the lead dummy should have a coefficient of zero. Otherwise, if there is reverse causality or some other type of endogeneity, then the coefficient will be different from zero. Regression 5.1 in Table 5 shows that the coefficient on the lead dummy of 0.004 (0.067) is extremely small and statistically indistinguishable from zero.

Therefore, although state divorce propensities do appear to be related to whether a state adopted unilateral divorce (and what type if adopted), the evidence bolsters the identifying assumption that prevailing state characteristics did not influence the timing of the legal change. These results support the use of panel data to identify the impact of unilateral divorce from variation in divorce laws within states over time, instead of across states at a point in time.

---

[17] All but five states had some type of unilateral divorce by the end of the time period, which does not yield a very informative correlation coefficient [only 0.09 (0.56)] with the divorce rate in 1968. The correlation coefficient rises when considering whether states adopted stricter versions of unilateral divorce. It is 0.27 (0.06) when unilateral divorce excludes separation requirements and 0.31 (0.03) when unilateral divorce excludes fault grounds for property settlement. The magnitudes are almost the same when excluding states already with unilateral divorce in 1968, and focusing on whether states without unilateral divorce adopted it.

[18] The correlation coefficient between the 1968 divorce rate and the subsequent year in which a state adopted unilateral divorce is −0.10 (0.56) when any type of unilateral divorce is considered and is of a similar magnitude and lack of significance when stricter types of unilateral divorce are considered.

B. *Timing*

Given the emphasis on the dynamics of the unobserved factors affecting divorce, it is instructive to explore whether the estimation properly captures the dynamics of *how* adopting a unilateral divorce law affected divorce. Two particular issues arise in this regard. One issue is whether the response to the law change was permanent, as modeled, or in fact transitory. Another is whether the length of the panel is appropriate.

First, the response to the legal change might have been a temporary boom in divorce—perhaps from a "backlog" of bad marriages breaking up earlier than otherwise—rather than a permanent change in divorce propensities. Adding dummies for the years of and immediately following the law change indicate the duration of *unilateral*'s impact. If the law change has only a temporary effect, we would expect a positive coefficient on the lag dummies and an insignificant coefficient on *unilateral.* The results are shown in regressions 5.2–5.4 of Table 5. In regression 5.4 with dummy variables for the year of and two years following the law change, the more general dynamic structure actually strengthens the long-run effect of *unilateral,* although the estimate of 0.536 (0.065) is not statistically different from the estimate in regression 3.4. The lag dummies have negative coefficients, revealing that the long-run effect was attained gradually. In the year of the law change, the impact was to raise the divorce rate by 0.419 [0.536–0.117], and in the following year by 0.446 [0.536–0.090]. The results demonstrate another reason why the estimate of *unilateral*'s impact was attenuated using Peters' and Allen's data—it was a cross section in the late 1970's, during the transition from mutual to unilateral divorce in many states.

Another issue involves the appropriate length of the panel. A long panel is useful, in principle, for capturing the background changes in divorce patterns over time. However, if the linear, or even quadratic, state trends are misspecified, then a longer panel could lead to a worse instead of better fit. If, further, the changing state divorce propensities are correlated with the adoption of unilateral divorce, then the longer is the panel and the worse the fit of the state trends, the more biased might become the estimate of *unilateral.* Most states switched to unilateral divorce in the 1970's, so the question that arises is whether the extra observations in the 1980's add useful information to the specification or worsen the estimate of *unilateral.*

To see whether the extra years of data alter the estimates, regression 5.5 in Table 5 cuts off the sample at 1979—still encompassing more than nine-tenths of the law changes. The estimate of *unilateral*'s impact in regression 5.5 is only slightly different, at 0.438 (0.055) compared to 0.447 (0.050) in regression 3.4, with a higher standard error. Adding one year at a time to the sample period from 1980 on induces relatively small, though not trivial, changes in the coefficient estimates, which range from 0.421 to 0.485. A partial explanation for this pattern arises from the evidence just noted of a gradual response of divorce rates in the remaining states switching to unilateral divorce. Adding years of data also improves the F-statistic testing the joint significance of the state trends. Therefore, the estimates do not indicate that the extra years of data, in which few states actually adopted unilateral divorce, distort the estimate of *unilateral,* but instead that they help in identifying the underlying trends in state divorce propensities.[19]

C. *Additional Features of the Estimation*

The estimated coefficients on the fixed effects and trends are of interest in themselves because they generate most of the explanatory power in the regressions. The coefficients on the year effects appear in Table 6. The year effects, which control for unobserved factors that affect divorce nationally, are significant and quite large, although adding state trends

---

[19] Since the estimates do not provide any conclusive evidence about when to cut off the panel, I chose to use data up to the latest year available.

TABLE 6—YEAR EFFECTS FROM REGRESSIONS IN TABLE 3

Independent variables:	Dependent variable: Divorce rate (divorces per 1,000 people)				
	3.1	3.2	3.3	3.4	3.5
Intercept	3.794 (0.057)	2.866 (0.195)	3.207 (0.175)	3.162 (0.222)	3.251 (0.276)
1969 year effect	—	0.225 (0.275)	0.236 (0.080)	0.221 (0.057)	0.196 (0.076)
1970	—	0.331 (0.276)	0.559 (0.081)	0.483 (0.066)	0.437 (0.123)
1971	—	0.510 (0.274)	0.775 (0.080)	0.682 (0.078)	0.614 (0.169)
1972	—	0.643 (0.274)	1.089 (0.081)	0.931 (0.093)	0.850 (0.211)
1973	—	0.837 (0.276)	1.365 (0.082)	1.163 (0.109)	1.062 (0.247)
1974	—	0.943 (0.275)	1.610 (0.083)	1.354 (0.126)	1.243 (0.279)
1975	—	1.232 (0.275)	1.840 (0.083)	1.574 (0.143)	1.454 (0.305)
1976	—	1.324 (0.273)	1.972 (0.083)	1.683 (0.160)	1.558 (0.326)
1977	—	1.269 (0.273)	1.922 (0.083)	1.623 (0.177)	1.496 (0.341)
1978	—	1.411 (0.271)	2.048 (0.083)	1.741 (0.194)	1.603 (0.352)
1979	—	1.571 (0.271)	2.211 (0.083)	1.895 (0.212)	1.760 (0.358)
1980	—	1.540 (0.270)	2.189 (0.082)	1.866 (0.230)	1.733 (0.360)
1981	—	1.558 (0.272)	2.217 (0.083)	1.889 (0.247)	1.765 (0.358)
1982	—	1.286 (0.270)	1.974 (0.082)	1.639 (0.265)	1.524 (0.353)
1983	—	1.202 (0.269)	1.873 (0.082)	1.528 (0.283)	1.423 (0.345)
1984	—	1.207 (0.272)	1.906 (0.083)	1.481 (0.301)	1.394 (0.336)
1985	—	1.200 (0.272)	1.913 (0.083)	1.476 (0.319)	1.416 (0.328)
1986	—	1.197 (0.270)	1.823 (0.083)	1.386 (0.337)	1.345 (0.323)
1987	—	1.105 (0.268)	1.691 (0.082)	1.337 (0.355)	1.319 (0.323)
1988	—	1.158 (0.280)	1.728 (0.085)	1.227 (0.373)	1.220 (0.330)

in regressions 3.4 and 3.5 reduces their magnitude a little and raises their standard errors. The pattern of the year effects mirrors the divorce trends in Figure 1, with the coefficients growing bigger through the 1970's, peaking around 1980, and then stabilizing. Thus, whatever were the factors common across the nation that affected divorce—more liberal attitudes towards divorce, for example—they increasingly raised divorce rates until 1980, which their influence steadied.

The coefficients on the state effects and trends from regression 3.4, shown in Table 7, tell us about geographical patterns in divorce propensities. Large negative intercepts are prevalent in the northeastern and north central states, so the propensity of their residents to divorce was the lowest at the beginning of the sample. But while starting low, the propensity to divorce rose somewhat in parts of the region, indicated by the positive coefficients on the slopes. Residents of southern states

TABLE 7—COEFFICIENTS ON STATE FIXED EFFECTS AND TRENDS IN REGRESSION 3.4

State	Coefficient (standard error) Fixed effect	 Trend
NORTHEAST		
Maine	—	—
New Hampshire	−0.080 (0.333)	−0.001 (0.027)
Vermont	−1.481 (0.395)	0.086 (0.033)
Massachusetts	−1.741 (0.238)	−0.016 (0.020)
Rhode Island	−2.125 (0.317)	0.046 (0.027)
Connecticut	−1.552 (0.253)	0.018 (0.021)
MIDDLE ATLANTIC		
New York	−2.271 (0.225)	0.076 (0.019)
New Jersey	−2.156 (0.234)	0.067 (0.020)
Pennsylvania	−1.845 (0.228)	0.028 (0.019)
EAST NORTH CENTRAL		
Ohio	0.105 (0.230)	0.013 (0.020)
Indiana	1.412 (0.345)	−0.014 (0.022)
Illinois	−0.235 (0.229)	−0.014 (0.019)
Michigan	−0.276 (0.232)	−0.032 (0.020)
Wisconsin	−1.823 (0.243)	0.041 (0.021)
WEST NORTH CENTRAL		
Minnesota	−1.539 (0.247)	−0.003 (0.021)
Iowa	−1.428 (0.257)	−0.000 (0.022)
Missouri	0.319 (0.242)	0.010 (0.020)

State	Coefficient (standard Fixed effect	 Tre
WEST NORTH CENTRAL (*continued*)		
North Dakota	−2.396 (0.357)	[illegible] [illegible]
South Dakota	−1.572 (0.349)	0.024 (0.025)
Nebraska	−1.416 (0.285)	0.013 (0.024)
Kansas	−0.096 (0.266)	0.012 (0.022)
SOUTH ATLANTIC		
Delaware	−0.656 (0.368)	0.054 (0.031)
Maryland	−1.112 (0.246)	0.010 (0.021)
District of Columbia	−0.243 (0.349)	0.047 (0.032)
Virginia	−1.164 (0.241)	0.054 [illegible]
West Virginia	−0.666 (0.275)	0.066 (0.023)
North Carolina	−1.116 (0.239)	[illegible] [illegible]
South Carolina	−1.572 (0.257)	0.061 (0.023)
Georgia	0.451 (0.241)	0.011 [illegible]
Florida	2.238 (0.234)	−0.013 (0.019)
EAST SOUTH CENTRAL		
Kentucky	−0.863 (0.251)	0.051 (0.023)
Tennessee	0.633 (0.245)	0.078 (0.021)
Alabama	0.773 (0.249)	0.027 (0.024)
Mississippi	0.102 (0.266)	0.017 [illegible]

TABLE 7—*Continued.*

State	Coefficient (standard error) Fixed effect	Coefficient (standard error) Trend
WEST SOUTH CENTRAL		
Arkansas	1.827 (0.271)	0.017 (0.023)
Louisiana	−1.600 (0.329)	0.034 (0.030)
Oklahoma	2.150 (0.290)	0.016 (0.023)
Texas	1.043 (0.228)	−0.003 (0.019)
MOUNTAINS		
Montana	0.943 (0.342)	−0.031 (0.029)
Idaho	1.235 (0.336)	−0.003 (0.028)
Wyoming	2.214 (0.433)	0.006 (0.035)
Colorado	1.114 (0.263)	−0.013 (0.022)
New Mexico	0.973 (0.662)	0.146 (0.063)
MOUNTAINS (*continued*)		
Arizona	2.289 (0.327)	−0.026 (0.025)
Utah	0.076 (0.301)	0.027 (0.025)
Nevada	16.679 (0.368)	−0.453 (0.029)
PACIFIC		
Washington	1.905 (0.249)	−0.047 (0.021)
Oregon	1.285 (0.266)	−0.020 (0.022)
California	1.171 (0.227)	−0.072 (0.019)
Alaska	2.120 (0.444)	0.022 (0.035)
Hawaii	0.046 (0.329)	−0.039 (0.027)

generally began with a moderate propensity to divorce that rose over the period, indicated by intercept coefficients close to zero and positive slope coefficients. Residents of western states started with the greatest propensity to divorce, which moderated in some states.[20]

Lastly, given the wide variation in divorce trends across states, a final check of whether *unilateral* is capturing other effects involves adding other state-level covariates to the regression. They might affect divorce but might not be captured fully by the state fixed effects and trends. Additional variables corresponding to Peters' covariates—education, race, urbanization, income, employment—were extracted for women from the *Current Population Survey* (CPS), March Annual Demographic Files (U.S. Department of Commerce, Bureau of the Census), and aggregated by state each year using the CPS weights. Including the extra variables in the regression leaves the coefficient on *unilateral* virtually unchanged.[21] This confirms that the state fixed effects and trends are capturing the important influences on divorce.

---

[20] The intercept and slope coefficients on Nevada are remarkably large because Nevada is a divorce haven, with very short residency requirements. This would bias upward the coefficient on *unilateral* if a couple that would have divorced in Nevada now divorces at home—a story corroborated by Nevada's large negative slope coefficient. However, the very high divorce rate in Nevada is misleading because Nevada has an extremely small resident population. Divorces in Nevada have ranged between [illegible] and 15,000 since the late 1960's, while nationwide divorces hit one million in 1975 and kept rising. The percentage of nationwide divorces obtained in Nevada, while close to 2 percent in the 1960's, has been in the neighborhood of 1 percent since then. Those magnitudes will not have much of an impact in estimating the coefficient on *unilateral* with weighted least squares.

[21] The coefficient on *unilateral* changes at most at the fourth decimal place. None of the state-level variables are

IV. Conclusion

Using panel data sheds new light on the impact of the no-fault revolution on divorce rates. It shows that the way one controls for geographical heterogeneity affects the estimation considerably, as Peters and Allen already demonstrated. But further, it shows that controlling flexibly for unobserved attributes of state populations that are correlated both divorce behavior and the divorce crucial. This is made possible by using gitudinal divorce data. I assembled a panel state-level divorce rates, which measures tually every divorce in the United States the entire duration of the law changes. estimation reveals a strong influence of lateral divorce: divorce rates would have about 6 percent lower if states had adopted unilateral divorce, accounting for percent of the overall increase between and 1988. The current move to tightening vorce requirements can therefore be expect to lower the number of divorces. Howev the estimates also demonstrate that other tors in addition to unilateral and no-fault vorce had a great deal to do with the incre in divorces in the United States.

---

significant at the 90-percent confidence level. The joint F-statistics on the state fixed effects and trends decline slightly but remain significant at better than 99.9 percent. Before 1978, not all states are identified separately in the CPS, reducing the sample size by 274.

APPENDIX

Dummy variables were created to account for the following irregularities in the data on divorce rates:

State	Time period	Reason
Kentucky	1978–1984	incomplete
Mississippi	1979, 1984–1985	incomplete
Arkansas	1973–1974, 1980, 1982–1985	incomplete
New Mexico	1968–1973, 1981–1982, 1986–1987	incomplete
New Mexico	1981–1982, 1984–1985	divorce petitions filed
California	1984–1986	include legal separ
Ohio	1981–1985	incomplete
Indiana	1968–1969, 1974, 1977–1986	include divorce peti
Indiana	1970–1977	incomplete
Oklahoma	1970–1974	include divorce peti
Arizona	1968–1972	include divorce peti
Michigan	1988	incomplete
District of Columbia	1987–1988	valid divorce certific
Mississippi	1988	incomplete
New Mexico	1988	incomplete
California	1988	incomplete

Louisiana had incomplete data for the entire series, which is accounted for in the estimation the Louisiana state effect. The regression results are quite similar when the observations above are omitted.

REFERENCES

Allen, Douglas W. "Marriage and Divorce: Comment." *American Economic Review*, June 1992, *82*(3), pp. 679–85.

Becker, Gary S. *A treatise on the fam* Cambridge, MA: Harvard Unive Press, 1981.

Brinig, Margaret F. and Buckley, F. H. Fault Laws and At-Fault People." *Int* al *Review of Law and Economics*, 1998 (forthcoming).

Cherlin, Andrew J. *Marriage, divorce, remarriage*. Cambridge, MA: Harvard University Press, 1992.

Freed, Doris Jonas, and Foster, Henry H., Jr. "Divorce in the Fifty States: An Outline." *Family Law Quarterly*, Fall 1977, *11*(3), pp. 297–313.

———. "Divorce in the Fifty States: An Overview as of 1978." *Family Law Quarterly*, Spring 1979, *13*(1), pp. 105–28.

———. "Divorce in the Fifty States: An Overview." *Family Law Quarterly*, Winter 1981, *14*(4), pp. 229–47.

Freed, Doris Jonas and Walker, Timothy B. "Family Law in the Fifty States: An Overview." *Family Law Quarterly*, Winter 1990, *23*(4), pp. 495–608.

Gruber, Jonathan and Hanratty, Maria. "The Labor Market Effects of Introducing National Health Insurance: Evidence from Canada." *Journal of Business and Economic Statistics*, April 1995, *13*(2), pp. 163–73.

Jacobson, Louis S.; LaLonde, Robert J. and Sullivan, Daniel G. "Earnings Losses of Displaced Workers." *American Economic Review*, September 1993, *83*(4), pp. 685–709.

Leland, John. "Tightening the Knot." *Newsweek*, February 19, 1996, *127*(8), pp. 72–73.

National Center for Health Statistics. *Vital statistics of the United States*, Vol. III. Washington, DC: U.S. Government Printing Office, 1960–1988.

Peters, H. Elizabeth. "Marriage and Divorce: Informational Constraints and Private Contracting." *American Economic Review*, June 1986, *76*(3), pp. 437–54.

———. "Marriage and Divorce: Reply." *American Economic Review*, June 1992, *82*(3), pp. 686–93.

Sepler, Harvey J. "Measuring the Effects of No-Fault Divorce Laws Across Fifty States: Quantifying a Zeitgeist." *Family Law Quarterly*, Spring 1981, *15*(1), pp. 65–102.

U.S. Department of Commerce, Bureau of the Census. *Current population survey*, March Annual Demographic Files. Washington, DC: U.S. Department of Commerce, Bureau of the Census, various years.

Divorce-Law Changes, Household Bargaining, and Married Women's Labor Supply

By JEFFREY S. GRAY *

The sharp rise in married women's labor-force participation rates over the past quarter century attests to the changing structure of the "typical" American family. The conventional economic model of household behavior, where the head maximizes family utility subject to total family resources, no longer seems appropriate. Cooperative bargaining models are the primary competing models of household decision-making (Marilyn Manser and Murray Brown, 1980; Marjorie B. McElroy and Mary Jean Horney, 1981). In these models each family member has input into the decision-making process, where the weight given each person's input depends upon his or her opportunities outside the family. Given an exogenous increase in a family member's income, cooperative bargaining models predict that family demand should shift to more reflect that person's preferences. In contrast, the pooled-income hypothesis of the conventional model implies that family demand is independent of which family member's income rises. These competing models of household behavior therefore yield different predictions regarding the impact of public policy targeted at individual family members.

Despite the methodological attractiveness of cooperative bargaining models, studies attempting to reject the conventional model's pooled-income assumption in favor of bargaining models have run into empirical obstacles. The primary difficulty is in obtaining data on individual income that is exogenous to family allocations of goods and time. Changes in individual income in a cross section, whether earned or unearned, will likely be endogenous with respect to family preferences or prices. Thus, the growing evidence rejecting the pooled-income hypothesis based on variation in family income (e.g., T. Paul Schultz, 1990; Duncan Thomas, 1990) should be interpreted cautiously. Significant changes in divorce laws during the 1970's, however, provide a ripe opportunity to test alternative models of household behavior avoiding the endogeneity problem plaguing earlier studies. The adoption of unilateral-divorce laws[1] has led to changes in the frequency and size of divorce settlements (Lenore Weitzman, 1985; Peters, 1986; Yoram Weiss and Robert J. Willis, 1993). Coupled with state laws governing marital property, a state's adoption of unilateral divorce alters how family assets are divided when a marriage ends. The potential redistribution of family wealth at divorce can be interpreted as an exogenous and unexpected shift in the unearned income of each spouse. Because alternative models of household decision-making generate different predictions regarding the impact this income shift will have on family time allocation, the actual time-allocation response to divorce-law changes enables the evaluation of the relative merit of these models.

---

* Welsh Consulting, 111 University Drive, East, Suite 205, College Station, TX 77840 (e-mail: jsg@econsulting.com). This paper is a condensed version of an earlier draft under the same title. I am grateful to Doug Simpson for programming assistance; to Karen Folk, Donald Deere, Bill Gale, Shoshana Grossbard-Shechtman, Michel Vanderhart, Carl Nelson, Robert Schoeni, the late Paul Taubman, and an anonymous referee for helpful discussions or comments; and especially to Andrew Foster and David Neumark for many useful discussions.

[1] Unilateral-divorce laws are a subset of no-fault divorce laws under which either spouse can obtain a divorce without the consent of his or her partner. Some states adopted provisions that effectively require each spouse's consent to obtain a no-fault divorce. Following Elizabeth Peters (1986), I classify states as having either "unilateral-" or "mutual-consent" divorce laws. This classification scheme is discussed later in the paper.

This paper tests the predictions made by alternative models of household behavior using three data sources: the *Current Population Survey* (CPS), the *1/1000 Public Use Micro Sample of the United States Census of Population and Housing* (Census), and the *Panel Study of Income Dynamics* (PSID). Two years from each of these data sets are used, one year before any state had adopted unilateral divorce, and a second year when over half the states had implemented the statutes. Women residing in states that did not adopt any changes to their divorce laws serve as a control group for examining the behavioral responses of those women whose states did adopt unilateral-divorce laws. This approach allows the impact of law changes to be examined in the absence of time trend. Using data at two different points in time also allows the test of how the adoption of unilateral-divorce laws contributes to the *growth* of married women's labor-force participation. It is possible that states with high married women's labor-force participation were the ones that adopted unilateral-divorce statutes. If so, then a cross-sectional analysis would find that married women's labor-force participation rates are higher in states with unilateral-divorce laws, even if labor-force participation rates did not change following the implementation of the laws. Therefore the estimation methods used in this paper have an advantage over cross-sectional studies such as Peters (1986) and Allen M. Parkman (1992), both of which find that residence in a unilateral-divorce state has a positive effect on whether a married woman participates in the labor force.

In contrast, these data suggest that unilateral divorce has no significant impact on married women's labor-force participation unless the underlying marital-property laws in each state are considered. Marital-property laws play an important role in determining how unilateral divorce affects each spouse's bargaining position. Nonetheless, previous studies of the impact of divorce laws give little, if any, attention to marital-property laws. Once these property laws are controlled for, I find that the labor-supply behavior of wives does appear to respond to their states adopting unilateral-divorce statutes. Data from the CPS and Census suggest that a wife's labor supply is an *increasing* function of her bargaining position within marriage.

The apparent responsiveness of married women to divorce laws that effectively redistribute property rights within marriage violates the neoclassical model's pooled-income hypothesis. The result is consistent with a bargaining model that acknowledges the importance of a wife's time spent working at home. Data from the PSID reinforce the importance of home production in household modeling, as the changes in the number of hours worked by married women in the labor market are due largely to opposing shifts in their home-production hours. The longitudinal aspect of the PSID allows us to examine the impact of divorce laws on the behavior of women who remain married throughout the sample period. These results suggest that the estimated changes in married women's work responses using pooled cross-sectional data may actually understate their true responses.

The next section discusses the role changing divorce and property laws have in theoretical models of household decision-making. Section II discusses the data used in the empirical investigation, and describes the classification of states with respect to their laws governing divorce. Section III presents the empirical methods used to test the impact of divorce-law changes, and reports the findings. Brief concluding remarks are contained in Section IV.

I. Bargaining Models and Divorce Laws

Unilateral-divorce laws allow either spouse to end a marriage simply by claiming that it has broken down. States' adoption of these laws during the 1970's placed many spouses, who were otherwise protected by mutual-consent-based statutes, in vulnerable positions. Which, if any, spouse was made more vulnerable depended on state laws governing the treatment of marital property at divorce. Equitable-distribution property laws, in principle, protected vulnerable parties by allowing courts discretion in redistributing household assets. However, courts did not have the same discretion in common-law and community-property jurisdictions.

Insofar as household assets were disproportionately held in the husband's name, the adoption of unilateral-divorce laws led to an implicit redistribution of assets towards the husband in common-law jurisdictions and towards the wife in community-property jurisdictions.

The distribution of assets within a household is irrelevant to the conventional neoclassical model of household behavior. Instead, the household head pools all family resources when determining the family's optimal behavior. Because intrafamily transfers of asset ownership have no effect on household allocations and decisions, the neoclassical model predicts that unilateral divorce should only affect the time allocations of wives through its effect on divorce probabilities. If unilateral divorce has no impact on divorce probabilities, then it should not affect married women's labor-supply behavior.

In contrast, cooperative bargaining models suggest that household behavior is sensitive to the distribution of wealth and earning power within the family. Bargaining models differ from their neoclassical counterpart by assuming that husbands and wives cooperatively bargain over possible decisions to be made by the household. The leverage each spouse has in the bargaining process is what he or she can attain by electing to dissolve the marriage and make decisions as an individual. Changes to divorce laws that affect the distribution of marital property at divorce alter each spouse's leverage in the bargaining process, and the predictions made by these models. In particular, bargaining models predict that a divorce-law change that increases (decreases) a wife's bargaining position will cause each spouse's commodity and time use to shift to more (less) strongly reflect her preferences. A shift in bargaining power is therefore likely to alter married women's demand for leisure, where "leisure" denotes all time not spent working in the labor market (see McElroy and Horney, 1981). How leisure is partitioned into pure leisure and home-production activities is important in interpreting married women's actual labor-supply response to changes in their bargaining position. Married women's time allocation response to changes in their bargaining positions is ultimately an empirical question that sheds light on the proper modeling of individual preferences and household behavior. The remainder of the paper examines empirically the effect of divorce and property laws on marital stability and married women's work decisions.

II. Data and Divorce-Law Classification

For the empirical analysis, states are categorized by the type of marital-property system they have, and whether they adopted unilateral divorce between 1970 and 1974. Five states adopted these statutes between 1975 and 1977. Because the evidence of bargaining by spouses in the form of labor-supply changes may take place over time, women in these states may not have yet responded to the adoption of unilateral divorce by the 1980 survey and are deleted from the sample. The primary analysis is also performed with these states included, and reported in the Appendix. The findings suggest that the omission, or inclusion, of these states does not alter the results qualitatively.

Most states adopting unilateral divorce completely abandoned fault-based statutes. Some states did retain the option for fault divorce, but in practice no-fault grounds are generally filed. States that do not qualify as having adopted unilateral-divorce laws are defined as having mutual-consent divorce laws. This category includes jurisdictions with only fault grounds for divorce, those that explicitly require mutual consent, and states where unilateral divorce is only possible after an extended period of separation. These states where unilateral divorce is possible, but with a large cost, constitute the borderline case. For many couples the cost of a lengthy separation may make a mutual-consent divorce more attractive. I follow Peters (1986) and do not define them as unilateral-divorce states.[2] Table 1 lists

---

[2] The analysis is replicated with this group defined as adopting unilateral-divorce laws, and with these states defined as adopting "unilateral-divorce-after-separation" statutes. The results from each of these specifications are reported in Appendix Table A2. The results from these alternative specifications suggest the labor-supply behavior of wives in these borderline states followed the trend of married women's labor supply in mutual-consent divorce states.

TABLE 1—CLASSIFICATION OF STATES BY DIVORCE LAWS

	Marital-property laws		
Common law	Equitable distribution		Community property
Unilateral divorce:			
Florida	Minnesota	Oklahoma	California
Georgia	Indiana	North Dakota	*Washington*
Alabama	Kentucky	Nebraska	*Arizona*
Massachusetts	Oregon	Kansas	*New Mexico*
Montana	Iowa	Michigan	*Nevada*
	Maine	*Delaware*	
	Wisconsin	*Wyoming*	
	Colorado		
Mutual-consent divorce:			
Maryland	District of Columbia		Louisana
Pennsylvania	New Jersey		Texas
New York	Illinois		*Idaho*
North Carolina	Arkansas		
South Carolina	South Dakota		
Mississippi	Connecticut		
Tennessee	*New Hampshire*		
West Virginia	*Hawaii*		
Ohio	*Alaska*		
Missouri	*Utah*		
Rhode Island	*Vermont*		
Virginia			

[illegible] States listed in italics were deleted from the CPS sample due to grouped coding by the CPS. Unilateral-divorce states are those states that adopted "incompatibility" and "irretrievable breakdown" grounds for divorce between 1970 and 1975, with the exceptions of Massachusetts, Montana, Rhode Island, Wisconsin, and Wyoming, all states that adopted [illegible] did so before 1975. Mutual-consent states include those with only-fault grounds for divorce, states that require [illegible] agreement, and states that require a lengthy separation before a divorce is granted. Three states, Massachusetts, [illegible], and Missouri, switched from common law to equitable distribution during the sample period.
[illegible] Doris J. Freed and Henry H. Foster (1973, 1979, 1981); Harvey J. Sepler (1981).

states that implemented unilateral divorce during the sample period, and categorizes states by their marital-property laws.[3]

The primary data used is the *1/1000 Public Use Micro Sample of the United States Census of Population and Housing* (Census) (U.S. Department of Commerce, Bureau of the Census). This data is a random sample of households from the entire Census performed every ten years. Observations from the 1960, 1970, and 1980 Censuses are used in the analysis. The primary sample is restricted to include married women ages 18 to 55 whose husbands are present. Sample means for variables used in the empirical analysis are reported in Appendix Table A1.

One limitation of the Census is that it contains limited information on an individual's hours worked. To analyze this component of labor supply, a second data source is used, the *Current Population Survey* (CPS), March Annual Demographic Files (U.S. Department of Commerce, Bureau of the Census). Two years of the CPS are used, 1968 and 1979. These years are chosen to provide a different time period for comparison with the Census results. Because the 1968 CPS groups smaller states together under a single code, the data does have limitations in analyzing the impact of state laws on individual behavior. States grouped together tend to be neighboring states such as Alabama and Mississippi. This grouping of states by the CPS will not pose a problem in estimation, as long as state specific heterogeneity among those states grouped together is not severe. However, if married women in states grouped together had different propensities to work prior to the divorce-law change, then the estimation would suffer from omitted variable bias. Comparing labor force participation estimates using the CPS to those using the Census provides insight into the magnitude of such biases. States grouped together having different marital-property laws were removed from the sample. Observations from states excluded from the [illegible] sample were similarly removed from the 197[illegible] sample. CPS sample means are reported in Appendix Table A1.

The final data source used is the cross-sectional sample from the *Panel Study of Income Dynamics* (PSID) (Survey Research Center). The empirical analysis uses data from the 1970 and 1980 surveys. Means from the PSID are presented in Table A1. The longitudinal aspect of the PSID allows me to examine how divorce laws affect the behavior of women who remained married throughout the sample period. Moreover, the PSID includes information on the number of hours worked by family members in household activities as well as in the labor market. This enables the direct test of how changes in bargaining positions affect wives' allocation of hours worked between the household and the market place, and their consumption of leisure. The limitations of the PSID include its relatively small sample size. Because behavioral changes at the state level are of interest, the limited number of observations in the PSID is particularly problematic. The PSID may provide insight into how divorce laws affect married women's time spent in home production, but these results should be interpreted with some caution.

III. Empirical Analysis

A. *Divorce Rates*

When mutual consent is necessary for a divorce, and there is symmetric information and trivial transaction costs, then a marriage will dissolve when the total utility achievable by each spouse if divorced exceeds the total [illegible] attainable by the partnership (see Gary Becker, 1981). The adoption of unilateral divorce laws by many states during the 19[illegible] lowered the necessary condition for divorce

---

[3] Five community-property states provide for "equitable distribution" of community property at dissolution (Arizona, Idaho, Nevada, Texas, and Washington). Because these states follow community-property laws for classifying assets, and because most of these states provide for a "substantially equal" division of property (W. S. McClanahan, 1982; Robert L. Mennell, 1982), they are classified as community-property states. The analysis is replicated with community-property jurisdictions that typically follow equitable-distribution guidelines defined as equitable-distribution states and reported in Appendix Table A2. The results are robust to this alternative classification scheme.

TABLE 2—DIVORCE LAWS' IMPACT ON DIVORCE PROBABILITIES

	(1) Differences-in-differences	(2) Probit	(3) Bivariate probit
Effect of change to unilateral divorce/equitable distribution	−0.002	0.004	0.005
	(−0.27)	(0.55)	(0.43)
Effect of change to unilateral divorce/common law	−0.0003	−0.006	0.020
	(−0.03)	(−0.64)	(1.21)
Effect of change to unilateral divorce/community property	0.006	−0.004	0.001
	(0.79)	(−0.63)	(0.04)

Notes: Asymptotic t-statistics in parentheses. The dependent variable is equal to one if the woman reports herself as divorced or separated from her husband at the time of the survey. Probit estimates reported in columns (2) and (3) are $\partial \Pr_i / \partial x_{ij} = \beta_j \cdot \phi(x_i'\beta)$, evaluated at the mean of x_i. In addition to divorce-law variables, each regression includes an intercept, age, years of schooling, children ever born and its square, child in household under the age of six, white, year, urban residence, and state dummy variables. The sample in columns (1) and (2) consists of 73,875 married and divorced or separated women ages 18 to 55 from the 1970 and 1980 data sets. The sample in column (3) includes an additional 11,671 never-married women ages 18 to 55 used to estimate the first-stage (selection) equation of the probability of ever-been married.

either spouse might seek to end the marriage when his or her utility if single would be greater than that received from being married. However, if the couple can easily make transfers within the marriage, the Coase theorem (R. H. Coase, 1960) implies that the frequency of divorce should be similar for couples facing either unilateral- or mutual-consent divorce laws.

A sample of unmarried, married, and divorced women ages 18 to 55 from the 1970 and 1980 Censuses is used to empirically test the Coase theorem's implication that unilateral-divorce laws should have no direct impact on marital instability.[4] Table 2 presents estimates of the impact of the change to unilateral divorce on a woman's likelihood of being divorced in each of the three marital-property regimes. The specification includes socioeconomic variables that are believed to affect divorce, as well as state dummy variables and variables controlling for divorce laws. These divorce-law variables consist of interactions between the unilateral-divorce variable and each of the marital-property variables. Each interaction term captures the change in the likelihood a woman in that marital-property jurisdiction is divorced following her state's adoption of unilateral divorce relative to the change in divorce probabilities of women residing in states that did not undergo any divorce-law change. Difference-in-difference coefficients obtained by estimating a linear probability model are reported in the first column in Table 2. These estimates reveal that unilateral divorce has little effect on divorce probabilities. As shown in column (2), estimation of a simple probit model yields similar results.

Because divorce is only an option for those who are married, the selection into marriage should also be controlled for when estimating divorce laws' impact on the probability of divorce. This selection is taken into account by estimating a bivariate probit model with sample selection.[5] Results from the second stage of this estimation procedure are presented in column (3). The coefficient associated with each divorce-law interaction term is low and insignificant, suggesting again that unilateral-divorce laws have little impact on state divorce rates.

---

[4] Here I am interested in how unilateral-divorce laws affect marital stability. Marital dissolution is defined to be [illegible] separation or divorce. In this paper I use the term [illegible] to mean marital dissolution. Means from the sam[illegible] divorced or separated women and never-married [illegible] reported in Appendix Table A1.

[5] The selection probit equation is the probability of ever having been married; the primary probit equation is the probability of being divorced. The estimator used is that proposed by Wynand P. M. M. van de Ven and Bernard M. S. van Praag (1981).

These results that unilateral-divorce laws did not significantly affect marital stability are consistent with studies using both retrospective data (Peters, 1986) and panel data (William R. Johnson and Jonathan Skinner, 1986). If the adoption of unilateral divorce alters the bargaining position of couples without increasing their likelihood of divorce, then transfers may have been made between spouses to prevent marital dissolution. One form of transfer could be a change in labor-force attachment.

B. *Labor Supply*

Previous studies examining the impact unilateral-divorce laws have on married women's labor supply give little attention to marital-property laws. Furthermore, each of these studies uses a single cross-sectional data set for its analysis.[6] It is possible that states with high married women's labor-force participation rates tend to adopt unilateral-divorce laws. If so, then an analysis that uses a single cross section could yield misleading results. This point is illustrated with the estimation of wives' labor-force participation. Column (1) in the first panel of Table 3 reports maximum likelihood probit estimates using cross-sectional data from 1980. The data suggest the probability a wife works in the paid labor force increases by 1.6 percent with residence in a unilateral-divorce state. This finding is similar to estimates found by Peters (1986) and Parkman (1992).

Using data at two points in time enables the estimation of how unilateral-divorce laws affect the *change* in wives' labor-force participation behavior. Column (2) presents probit estimates using data from both 1970 and 1980. State dummy variables are added as regressors to control for state-specific effects contribu married women's labor-force participation cisions.[7] A year dummy variable is includ capture the increasing trend of married wome participation during the decade, as well as gregate business-cycle effects. In this speci tion the unilateral-divorce variable measure change in labor-force participation of wives states that adopted unilateral divorce relative the labor-force participation changes of wives other states. This coefficient is small and nificant, suggesting that without considering marital-property laws in each state, the adop of unilateral divorce had no significant imp on married women's increasing labor-force ticipation rates during the 1970's.

Estimating the impact of unilateral-div laws alone on married women's labor-f participation may obscure how these laws fect wives' behavior. Depending on state governing marital property, unilateral div may encourage or discourage women's la force participation. Column (3) presents mates of how married women's labor-fo participation behavior responds to shifts their bargaining positions. The interaction tween the unilateral-divorce and the com nity-property dummy variables represents labor-force participation of married wo changed in community-property states foll ing their state's implementation of unil divorce, relative to the change in labor-f participation of wives in states that did adopt unilateral divorce. As shown in fourth row, this estimate is positive and nificant at the 5-percent level, suggesting wives are more likely to increase their force participation, or less likely to ex

---

[6] Johnson and Skinner (1986) use a single cross section of the PSID when looking at the impact of divorce laws. Peters (1986) and Parkman (1992) each use the 1979 March\April CPS.

[7] While state dummy variables control for state heterogeneity, they do not control for state charac that influence *how* unilateral-divorce laws affect women's bargaining positions. For example, the ing position of a married woman residing in a st a very generous welfare system may be less affec the introduction of no-fault divorce than a wife in a state with a limited welfare system. Another im example of state attributes not controlled for by dummy variables are state laws governing mari erty. The role of these laws in unilateral divorce's on married women's bargaining positions and ply is explored in subsequent detail.

TABLE 3—DIVORCE LAWS' IMPACT ON WIVES' EMPLOYMENT AND HOURS WORKED

	Census			CPS	
	(1) Employed cross section	(2) Employed fixed effects	(3) Employed fixed effects	(4) Employed fixed effects	(5) Hours fixed effects
A. Cross section and fixed-effects specifications					
Unilateral divorce	0.016 (2.90)	0.006 (0.74)	—	—	—
Effect of change to unilateral divorce/equitable distribution	—	—	0.013 (1.26)	0.001 (0.10)	0.78 (1.54)
Effect of change to unilateral divorce/common law	—	—	−0.036 (−2.35)	−0.062 (−3.47)	−1.76 (−2.70)
Effect of change to unilateral divorce/community property	—	—	0.023 (1.96)	0.048 (3.07)	2.47 (4.08)
Sample size	34,069	63,615	63,615	38,984	38,984

	(1) State interactions	(2) Fully interacted model
B. Controlling for 1960 to 1980 trend		
Effect of change to unilateral divorce/ equitable distribution	−0.014 (−0.75)	−0.014 (−0.71)
Effect of change to unilateral divorce/ common law	−0.049 (−1.73)	−0.053 (−1.84)
Effect of change to unilateral divorce/ community property	0.044 (2.03)	0.052 (2.37)
Sample size	93,210	93,210

Notes: Asymptotic t-statistics in parentheses. For estimation of whether employed, dependent variable equals one if the person worked in the preceding survey year; for hours estimation, dependent variable is the number of hours worked in the survey week. Probit estimates reported are $\partial P_i/\partial x_{ij} = \beta_j \cdot \phi(x_i'\beta)$; column (5) reports Tobit estimates as $\beta_j \cdot \Phi[x_i'\beta/\sigma]$, each evaluated at the mean of x_i. In addition to divorce-law variables, each regression includes an intercept, [illegible] income, husband's earnings, age, years of schooling, children ever born and its square, and child in household under the age of six, white, year, urban residence, and state dummy variables. Year and state dummy variables are excluded from the column (1) specification. State interactions with a trend variable are included in Panel B specifications. The fully interacted model in Panel B includes interaction terms between the independent variables and the trend variable. In Panel A, columns (1)–(3) use Census data and columns (4)–(5) use CPS data; column (1) only includes data from [illegible]. Panel B uses Census data from 1960, 1970, and 1980.

[illegible] force, when their bargaining position im[illegible] In contrast, the adoption of unilateral [illegible] in common-law states leads to a re[illegible] in married women's labor-force par[illegible], suggesting that wives are less likely to join the paid labor force, or are more likely to withdraw from the labor force, when their bargaining position worsens. Finally, how the adoption of unilateral-divorce laws affects the labor-force participation of wives in states with equitable-distribution property laws is captured by the interaction term between the unilateral-divorce and the equitable-distribution dummy variables. It is argued that women in these states did not experience a change in their bargaining position due to the implementation of unilateral divorce. This is evidenced in the data as unilateral-divorce laws have no significant impact on married women's labor-force behavior in these states.

If a shift in a wife's bargaining position affects her labor-force participation decision, it should similarly affect her hours-worked decision. The CPS is used to analyze the impact divorce laws have on the number of hours a woman chooses to work. Column (4) reports probit estimation of the impact unilateral-divorce laws have on married women's labor-force participation using the CPS data. The estimated coefficients have the same sign as those found using the Census data. The coefficients are larger in magnitude, which might result from the different subsample of states used due to coding by the CPS. Column (5) presents estimates of married women's hours-worked decisions. Tobit estimation is used to account for the truncation of hours worked at zero. Unilateral divorce has a similar effect on married women's hours-worked decisions as it does on their labor-force participation decisions.

Therefore, the CPS and the Census both provide evidence that a wife's labor supply is an increasing function of her bargaining position within marriage. However, it is possible that common-law states that adopted unilateral divorce during the early 1970's had a slower trend in married women's labor-force participation rates for many years, while community-property states experienced an extended increase in their married women's labor-force participation rates, independent of their implementation of divorce laws. If these corresponding trends in labor-force participation rates preceded the divorce law's implementation, then it would be inappropriate to ascribe changes in labor-force participation rates during the 1970's to the adoption of the laws. Using Census data from 1960, as well as 1970 and 1980, we can control for state trends in the labor-force behavior of married women.

The second panel in Table 3 presents probit estimates of married women's labor-force participation combining data from each
A trend variable is interacted with e
variable to control for state-specific
married women's labor-force parti
Results using this specification are
in column (1). When controlling for
in participation rates, the estimated i
the adoption of unilateral-divorce
married women's labor-force partici
community-property and common-l
increases in magnitude and remains si
at the 10-percent level. Column (2)
results when the model is fully intera
the trend variable. This specificatio
each independent variable to have a
fect. The results using this specific
similar to those reported in column (1
fore, the observed changes in
women's labor-force participation f
the adoption of unilateral divorce c
explained by different longer run stat

C. *Home Production and Leis*

Whether the observed changes in
women's labor supply results from ch
their leisure consumed or their home
tion hours has important implication
proper modeling of household behavi
4 presents estimates of how divorce la
the allocation of a wife's time amon
market work, home production, and
leisure activities. The determinants of
time spent in the labor market and
production are estimated jointly u
expectation-maximization algorithm
Dempster et al. (1977). The determi
leisure are then calculated such that
of the coefficients of each variable in
tem equals zero (see McElroy, 199
first three columns use a sample of
women from the 1970 and 1980 surve
PSID.[8] While the point estimates pre

---

[8] Due to the small sample size of the PSID, s
are unrepresented, while others have very fe
tions. As a result, each state's married wom
force participation rate in 1970 together wi
division dummy variables are included to co
gional heterogeneity in married women's pr
work in the labor market and at home.

TABLE 4—DIVORCE LAWS' IMPACT ON MARRIED WOMEN'S TIME ALLOCATIONS

	Women married in either 1970 or 1980			Women married throughout sample period		
	Labor market	Home hours	Pure leisure	Labor market	Home hours	Pure leisure
Effect of change to unilateral divorce/equitable distribution	−1.40 (−1.14)	1.12 (0.91)	0.28 (0.64)	−2.44 (−1.55)	1.64 (1.00)	0.80 (0.89)
Effect of change to unilateral divorce/common law	−1.91 (−0.88)	2.94 (1.36)	−1.03 (−0.15)	−7.19 (−2.31)	5.16 (1.61)	2.03 (0.74)
Effect of change to unilateral divorce/community property	1.34 (0.74)	−4.28 (−2.32)	2.94 (1.70)	0.62 (0.25)	−6.85 (−2.61)	6.23 (2.33)
Sample size	2,581	—	—	1,336	—	—

Notes: Asymptotic *t*-statistics in parentheses. See footnote to Table 3 for other independent variables included; Census region dummy variables are included rather than state dummy variables. Tobit estimates reported as $\beta_j \cdot \Phi[(x_i'\beta)/\sigma]$, evaluated at the mean of x_i. Sample used in first three columns consists of a pooled cross section of married women in 1970 and 1980; the sample used in last three columns is restricted to women who remained married between 1970 and 1980.

the first column are consistent with the CPS results that a wife's labor supply is an increasing function of her bargaining position within marriage, none of the coefficients is statistically significant. The second column reveals that following their states' implementation of unilateral divorce, wives residing in community-property states decreased their time spent in home production and wives living in common-law states (insignificantly) increased their home-production hours, relative to the change in home-production time of other wives. Thus in contrast to labor-market hours, a wife's hours at home is a decreasing function of her bargaining power. The net result of these changes in time allocation is consistent with the standard bargaining model as leisure consumption decreases in states where wives lose bargaining power, and increases in states where wives' bargaining power increases.[9]

These results suggest that the changes in labor-market activity found using the Census and the CPS is likely due to wives changing their commitment to home production.

These findings have implications for the nature of household preferences over a wife's nonmarket time before the divorce-law changes. Assuming the husband and wife's utility functions are monotonically increasing in her leisure and in her home-production time, the results are consistent with the standard cooperative bargaining model when the wife places a relatively higher value on an additional hour of her leisure than her husband, but he receives relatively greater utility from an additional hour of her home-production time than she does. Given these preferences, the cooperative bargaining model predicts that an increase in her bargaining position will lead to a decrease in her home-production time and an increase in her time spent in leisure. The empirical results using the Census and the CPS are consistent with these predictions, where the magnitude of a wife's change in her home production time following the divorce-law change dominates the magnitude of her change in leisure hours. The net result is the observed finding that a wife's labor supply is an increasing function of her bargaining position within marriage.

---

[9] Measurement error associated with home-production hours could contribute to the imprecision of the [illegible] estimates. If measured home hours is not ex-[illegible] of all home-production activities, then the leisure [illegible] would also reflect these unobserved tasks, [illegible] estimated impact of divorce laws on pure leisure in common-law and community-property states [illegible] biased towards zero.

The empirical analysis so far has been restricted to using pooled cross sections of married women. These samples include women who married following the adoption of the divorce laws, so prenuptial contracts could be drawn to mitigate the impact of the laws. Moreover, the adoption of unilateral-divorce statutes may alter the likelihood certain types of women enter or exit marriage in each marital-property regime. Unfortunately the data do not offer any compelling identifying variables that would permit the estimation of a joint model of marriage, divorce, and time allocation. An alternative approach to address selection into and out of marriage is to compare the results found above to those found when restricting the sample to women who remain married throughout the time period that the laws were introduced. Results from this alternative approach are reported in the last three columns of Table 4. With the exception of married women's labor-market response to an increase in their bargaining power, the estimates increase in magnitude and significance when using the subsample of women who remain married throughout the sample period. This indicates that selection may be taking place into and out of marriage that mitigates the observed behavioral responses when using pooled cross sections of married women.[10] Due to the relative changes in the sizes of the coefficients, the estimated impact of a decrease in a wife's bargaining position on her leisure consumption becomes positive. Thus, when their threat point decreases, it is sufficient for most wives to shift their allocation of work
hours from the labor market to home pr
tion, without decreasing their consumpti
leisure. This suggests that husbands pl
high value on their wives' time spent wo
at home.

IV. Conclusion

This paper examined married women
sponses to the reassignment of property r
within marriage brought about by n
changes in divorce laws. Three separate
sources were used to obtain three prin
findings. First, consistent with the Coase
orem, the reassignment of property ri
within marriage does not lead to an inc
in the likelihood of divorce. Second, w
who were favored by this reassignmen
property rights increased their labor su
and wives who were disadvantaged decr
their labor supply, relative to married w
who did not face a redistribution of m
property rights. Third, these changes in w
hours of market work primarily reflect ch
in their home-production hours rather
changes in their consumption of leisure
second finding provides explicit eviden
jecting the neoclassical model's assumpti
income pooling within the household.
finding also appears inconsistent with the
dard bargaining model of household beh
if leisure is defined as nonlabor market
However, the third finding illustrates th
empirical results are consistent with the
dard bargaining model, so long as nonm
time is divided between leisure and home
duction time.

While the empirical results stress th
portance of home production in mod
household behavior, they also have direc
icy implications. In an attempt to impro
economic position of divorced women, se
states are considering replacing their e
ble-distribution property laws with co
nity-property laws. The findings in this
suggest that these proposed property
changes may provide additional incentive
wives to increase their labor supply and re
their home-production time. Moreover
responsiveness of wives to changes in
bargaining positions within marriage ha

---

[10] An earlier version of this paper, Gray (1994), uses the Census sample to determine whether unilateral-divorce laws led to any selection taking place into and out of marriage along unobservable characteristics. This selection is identified via estimated interaction effects of divorce laws with age, race, and education on marriage, divorce, and labor-force participation probabilities. The findings suggest that the adoption of unilateral divorce may have led to a selection of high-wage women out of marriage in states with community-property laws. Less selection appears to have taken place by married women out of common-law states. Tables summarizing these results are available from the author upon request.

[illegible] welfare implications for divorce-law [illegible]. Any divorce-law change that alters the financial well-being of divorcing women and their children will also impact the welfare of individuals in families that do not dissolve. These indirect effects should not be ignored when designing effective social and economic policies.

APPENDIX

TABLE A1—SAMPLE MEANS

Variable	Married, spouse present			Divorced or separated		Never married	
	(1960)	(1970)	(1980)	(1970)	(1980)	(1970)	(1980)
	A. Means from Census						
Age	37.25	37.11	36.72	37.28	36.25	25.79	24.97
Years of schooling	10.66	11.48	12.25	10.26	11.92	11.33	12.63
Child less than six	0.42	0.35	0.30	0.27	0.16	0.12	0.09
Children born	2.39	2.51	2.27	2.47	2.19	0.23	0.29
White	0.89	0.89	0.84	0.70	0.69	0.79	0.70
Urban residence	0.71	0.73	0.80	0.85	0.85	0.81	0.86
Unilateral divorce	0.00	0.00	0.43	0.00	0.45	0.00	0.41
Common law	0.44	0.44	0.40	0.43	0.41	0.47	0.42
Community property	0.21	0.22	0.24	0.26	0.28	0.21	0.24
Equitable distribution	0.35	0.34	0.36	0.31	0.32	0.32	0.34
Husband's income ($000)	17.42	23.14	23.99	—	—	—	—
Husband's earnings ($000)	12.61	18.09	18.58	—	—	—	—
Employed	0.42	0.53	0.64	—	—	—	—
Sample size	29,595	29,595	34,069	3,372	6,888	6,794	11,277

[illegible]	Means from CPS		Means from PSID			
	Cross sections		Cross sections		Married throughout	
	(1968)	(1979)	(1970)	(1980)	(1970)	(1980)
	B. Means from CPS and PSID					
[illegible]	0.54	0.65	0.58	0.71	0.57	0.68
[illegible]	13.61	22.37	13.81	18.42	12.41	17.55
[illegible]	—	—	33.75	28.24	35.77	30.09

TABLE A1—*Continued.*

	Means from CPS		Means from PSID		
	Cross sections		Cross sections		Married
Variable	(1968)	(1979)	(1970)	(1980)	(1970)
B. Means from CPS and PSID—*Continued*					
Leisure hours	—	—	120.46	121.35	119.83
Age	37.42	36.53	36.52	36.69	32.72
Years of schooling	12.33	13.06	4.34	4.63	4.50
Child under six	0.36	0.31	0.63	0.61	0.61
Children	1.95	1.64	1.76	1.33	2.11
White	0.91	0.91	0.89	0.89	0.91
Nonlabor income ($000)	0.83	1.49	2.97	4.71	2.35
Husband earnings ($000)	12.95	13.78	18.77	19.39	20.10
Urban	0.65	0.69	0.63	0.63	0.64
Unilateral divorce	0.00	0.40	0.00	0.41	0.00
Common law	0.50	0.42	0.44	0.41	0.45
Community property	0.22	0.23	0.17	0.16	0.15
Equitable distribution	0.28	0.35	0.39	0.44	0.40
Sample size	20,259	18,725	1,363	1,218	668

Notes: Each sample contains women ages 18 to 55. The three data sources have different definitions for some (such as years of schooling, urban residence, and nonlabor income). However, the definitions are consistent sample.

TABLE A2—DIVORCE LAWS' IMPACT ON WIVES' EMPLOYMENT USING ALTERNATIVE CLASSIFICATION

Effect of divorce-law change to:	(1)	(2)	(3)	(4)
Unilateral divorce/equitable distribution	0.012 (1.21)	0.010 (1.07)	0.016 (1.51)	0.015 (1.40)
Unilateral divorce/common law	−0.033 (−2.36)	0.038 (−2.45)	−0.015 (−1.16)	−0.035 (−2.23)
Unilateral divorce/community property	0.021 (1.89)	0.022 (1.85)	0.020 (1.76)	0.025 (2.09)
Unilateral divorce after separation/ equitable distribution	—	—	—	0.022 (0.76)

TABLE A2—*Continued.*

Effect of divorce-law change to:	(1)	(2)	(3)	(4)	(5)
Unilateral divorce after separation/ common law	—	—	—	0.025 (1.15)	—
Unilateral divorce after separation/ community property	—	—	—	−0.018 (−0.61)	—
Sample size	63,615	67,439	63,615	63,615	63,615

Notes: See notes to Table 3 for regression details. Asymptotic *t*-statistics in parentheses. Probit estimates reported in columns (2)–(5) are $\partial P_i/\partial x_{ij} = \cdot\beta_j \cdot \phi(x_i'\beta)$, evaluated at the mean of x_i. Column (1) reports linear probability estimates of the baseline regression. States that adopted unilateral divorce between 1975 and 1980 are included in estimates reported in column (2). In column (3), states that allow unilateral divorce after a lengthy separation are defined as having adopted unilateral-divorce laws; these states are defined as adopting a distinguishable form of unilateral-divorce laws in the specification shown in column (4). Column (5) reports estimated coefficients when community-property states with essentially equitable-distribution statutes (Texas, Arizona, and Nevada) are classified as having equitable-distribution property laws.

REFERENCES

Becker, Gary S. *A treatise on the family*. Cambridge, MA: Harvard University Press, 1981.

Coase, R. H. "The Problem of Social Cost." *Journal of Law and Economics*, October 1960, *3*(1), pp. 1–44.

Dempster, A. P.; Laird, N. M. and Rubin, D. B. "Maximum Likelihood From Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, Series B, 1977, *39*(1), pp. 1–22.

Freed, Doris Jonas and Foster, Henry H. "Economic Effects of Divorce." *Family Law Quarterly*, Fall 1973, *7*(2), pp. 275–343.

———. "Divorce in the 50 States: An Overview as of 1978." *Family Law Quarterly*, Spring 1979, *13*(1), pp. 105–28.

———. "Divorce in the 50 States: An Overview." *Family Law Quarterly*, Winter 1981, *14*(3), pp. 229–84.

Gray, Jeffrey S. "Divorce Law Changes, Household Bargaining, and Married Women's Labor Supply." Mimeo, University of Illinois, 1994.

Johnson, William R. and Skinner, Jonathan. "Labor Supply and Marital Separation." *American Economic Review*, June 1986, *76*(3), pp. 455–69.

Manser, Marilyn and Brown, Murray. "Marriage and Household Decision-Making: A Bargaining Analysis." *International Economic Review*, February 1980, *21*(1), pp. 31–44.

McClanahan, W. S. *Community property law in the United States*. San Francisco, CA: Bancroft-Whitney, 1982.

McElroy, Marjorie B. "The Empirical Content of Nash-Bargained Household Behavior." *Journal of Human Resources*, Fall 1990, *25*(4), pp. 559–83.

McElroy, Marjorie B. and Horney, Mary Jean. "Nash-Bargained Household Decisions: Toward a Generalization of the Theory of Demand." *International Economic Review*, June 1981, *22*(2), pp. 333–49.

Mennell, Robert L. *Community property law: In a nutshell*. St. Paul, MN: West, 1982.

Panel Study of Income Dynamics. Survey Research Center, Economic Behavior Program, 1970, 1980.

Parkman, Allen M. "Unilateral Divorce and the Labor-Force Participation Rate of Married Women, Revisited." *American Economic Review*, June 1992, *82*(3), pp. 671–78.

Peters, H. Elizabeth. "Marriage and Divorce: Informational Constraints and Private Contracting." *American Economic Review*, June 1986, *76*(3), pp. 437–54.

Schultz, T. Paul. "Testing the Neoclassical Model of Family Labor Supply and Fertility." *Journal of Human Resources*, Fall 1990, *25*(4), pp. 599–634.

Sepler, Harvey J. "Measuring the Effects of No-Fault Divorce Laws Across 50 States: Quantifying a Zeitgeist." *Family Law Quarterly*, Spring 1981, *15*(1), pp. 65–102.

Thomas, Duncan. "Intra-household Resource Allocation: An Inferential Approach." *Journal of Human Resources*, Fall 1990, *25*(4), pp. 635–64.

U.S. Department of Commerce, Bureau of the Census. *1/1000 public use micro sample of the United States census of population and housing*. Washington, DC: U.S. Department of Commerce, Bureau of the Census, 1960, 1970, 1980.

———. *Current population survey,* March Annual Demographic Files. Washington, DC: U.S. Department of Commerce, B of the Census, 1968, 1979.

van de Ven, Wynand P. M. M. and van P Bernard M. S. "The Demand for Deduct in Private Health Insurance: A Probit M with Sample Selection." *Journal of Ec metrics*, November 1981, *17*(2), pp. 2 52.

Weiss, Yoram and Willis, Robert J. "Tran Among Divorced Couples: Evidence Interpretation." *Journal of Labor Eco ics*, October 1993, *11*(4), pp. 629–79.

Weitzman, Lenore. *The divorce revolution: unexpected social and economic con quences for women and children in A ica*. New York: Free Press, 1985.

Caps on Political Lobbying

By YEON-KOO CHE AND IAN L. GALE*

The cost of political campaigns in the United States has risen substantially in recent years. For example, real spending on congressional election campaigns doubled between 1976 and 1992 (Steven D. Levitt, 1995). There are many reasons why increased campaign spending might be socially harmful. First, increased spending means increased fund-raising, which may keep politicians from their legislative duties.[1] Second, a lobbyist who makes a large campaign contribution may have undue influence on electoral outcomes, on the shaping of legislation, or on the outcome of regulatory proceedings.[2] That is, the socially preferred candidate or legislation may not prevail. Likewise, a lobbyist involved in a regulatory matter or a competition for a government contract may benefit unduly from a legislator's intervention.[3] Third, a perception that campaign contributions purchase influence may lead to increased tolerance of corruption in the private sector.

A desire to control campaign spending has spawned many initiatives to limit both campaign contributions and spending, beginning with the passage of the Federal Election Campaign Act (FECA). Political action committees can contribute at most $5,000 per election to a candidate, while individuals can contribute at most $1,000. (Restrictions have also been put on in-kind contributions, making it more difficult to circumvent these limits.)[4] While direct restrictions on campaign spending have proven difficult to implement, recent initiatives aim to impose voluntary spending limits and stricter limits on contributions.[5]

Despite the existing legislation and the proposals to limit contributions, little is known about the impact of contribution limits on aggregate expenditures. While it is intuitively appealing that aggregate expenditures would drop, we challenge that intuition here. We study a lobbying game and show that a cap on individual lobbyists' expenditures may have the perverse effect of increasing aggregate expenditures and lowering total surplus. This result suggests that a cap on campaign contributions may increase aggregate contributions.[6]

---

* Che: Department of Economics, University of Wisconsin, Madison, WI 53706 and Yale Law School, New Haven, CT 06520; Gale: Department of Economics, Georgetown University, Washington, DC 20057. We thank three anonymous referees for numerous valuable comments.

[1] The following quotation makes this point: "Critics maintain that high campaign costs force candidates to devote an inordinate amount of time to raising money. They also hold that special interest groups seeking to exercise influence by satisfying the candidates' need for campaign funds threaten the integrity of the election and governmental processes." (Herbert E. Alexander and Monica Bauer, 1991 pp. 1–2.)

[2] The empirical evidence of a link between campaign contributions and roll-call votes in the House of Representatives and Senate is mixed (Levitt, 1995). There is evidence that lobbyists' influence is felt before legislation reaches the floor, however. Richard L. Hall and Frank W. Wayman (1990) examined committees of the House of Representatives, finding a significant relationship between campaign contributions and members' efforts to shape legislation at the committee stage. Thomas Romer and James M. Snyder, Jr. (1994) found a significant relationship between committee assignments and political action committee (PAC) contributions. In particular, they found that seniority is rewarded, suggesting that contributors target influential members. John R. Wright (1990) studied committee voting and found that campaign contributions [illegible] subsequent lobbying.

[3] In one legendary case, five senators met with officials of the Federal Home Loan Bank Board on behalf of a banker who had contributed $1.3 million to the senators and their parties (Alexander, 1991 pp. 116–17).

[4] Recent legislation restricted the types of gifts that members of Congress may accept (*Congressional Quarterly*, 1995a).

[5] Mandatory spending limits were struck down by the Supreme Court in *Buckley v. Valeo*, 424 U.S. 1 (1976). The FECA was then amended to incorporate public funding for presidential candidates who voluntarily accept spending limits. Some recent proposals include voluntary spending limits for congressional campaigns and stricter contribution limits (*Congressional Quarterly*, 1996).

[6] Lobbying organizations provide a large and growing fraction of total campaign contributions. For example,

The next section presents the model and describes the equilibrium when lobbyists are unconstrained. We then solve for the equilibrium when lobbyists face a cap on individual expenditures. When a cap constrains the high-valuation lobbyist, a lobbyist with a lower valuation for the political prize becomes relatively more aggressive. As a consequence, total lobbying expenditures may rise. Since the high-valuation lobbyist's probability of winning the prize drops, the cap reduces total surplus if private and social valuations coincide. Concluding remarks are contained in the final section.

I. The Model

Two risk-neutral lobbyists seek a political prize.[7] The prize could be a government contract, a military base, or a license to produce a good or service. An incumbent politician determines who will receive the prize. Ethics legislation prevents the open sale of political prizes, so the politician will award the prize to the lobbyist who spends more. We do not model the politician's objective function explicitly, but two interpretations of her behavior are possible.[8] First, the politician may be self-interested. A self-interested politician wishes to extract rents from the lobbyists. Although she cannot sell the prize openly, the politician may accept campaign contributions or in-kind contributions.[9] Second, the politician may be benevolent. In this case, the politician wishes to award the prize to the lobbyist who will add more to social welfare. A benevolent poli[illegible] will award the prize to the lobbyist who sp[illegible] more, if she does not know the individual [illegible] uations, since a lobbyist with a higher va[illegible] tion will spend more, on average.[10] (Re[illegible] legislation requires lobbyists to disclose [illegible] lobbying expenditures for each issue on w[illegible] they lobby members of Congress.)[11]

Gordon Tullock (1980) considered lob[illegible] ing in a setting where an individual's pro[illegible] bility of winning a political prize dep[illegible] directly on his lobbying expenditures.[12] T[illegible] case known as the "all-pay auction" has [illegible] analyzed by Arye L. Hillman and John [illegible] Riley (1989) and Michael R. Baye et [illegible] (1993, 1996). In an all-pay auction, bid[illegible] submit nonnegative bids simultaneously [illegible] the prize is awarded to the highest bidder. [illegible] novel feature is that all bidders pay their [illegible] which is appropriate here since a lobby[illegible] contributions are not typically returned if [illegible] efforts are unsuccessful. The all-pay auction [illegible] also appropriate for other rent-seeking [illegible] such as labor-market tournaments, as well [illegible] for research and development contests.

We analyze the all-pay auction when [illegible] ders face an exogenous cap on bids. In keep[illegible] with the all-pay terminology, we refer to [illegible] politician as "the seller" and to the lobby[illegible] as "the bidders." Bidder i's valuation of [illegible] prize is v_i, and $v_1 > v_2 > 0$.[13] If bidder i [illegible] with a bid b_i his payoff is $v_i - b_i$, where[illegible] payoff is $-b_i$ if he loses. (If they tie, the [illegible] ders are equally likely to win.) The rules of [illegible] game and the payoffs are known by the [illegible] ders, who maximize their individual exp[illegible] payoffs. We look for a Nash equilibrium [illegible]

PACs contributed nearly half of all money spent by incumbents in the 1992 elections for the House of Representatives (Levitt, 1995).

[7] Another interpretation of the model is that the lobbyists are candidates for political office, and the candidate who spends more will win.

[8] A complementary view is that lobbying is directly informative (see David Austen-Smith and Wright [1992] or Austen-Smith [1995]). The information might concern the impact on constituents of pending legislation, and it could take the form of technical research or public opinion surveys, for example.

[9] Even if the politician wished to hold a standard auction rather than an all-pay auction, it might be difficult to commit to return contributions from unsuccessful lobbyists, or for lobbyists to commit to make a contribution if awarded the prize.

[10] In Robert W. Helsley and Arthur O'Sullivan [illegible] contributions to a referendum campaign signal the [illegible] valuations of the competing lobbies, so citizens' [illegible] pend on which lobby contributes more to the refe[illegible] campaign.

[11] See *Congressional Quarterly*, 1995b.

[12] Tullock supposed that lobbyist i's probability [illegible] ning the political prize is $b_i^R/(\Sigma_{j=1}^{n} b_j^R)$, where b_j [illegible] amount lobbyist j spends and R is a constant. In [illegible] pay auction, $R = \infty$.

[13] In the case of a political prize that may [illegible] awarded until after the next election, the valuation [illegible] incorporate the probability that the politician [illegible] reelected.

bidding strategies. The bids accrue to the seller.

Absent a cap on bids, bidder 1 wins the prize with probability $1 - (v_2/2v_1) > 1/2$ in equilibrium and the seller's expected revenue is $(v_1 + v_2)/2v_1 < v_2$.[14] Bidder 2 will bid less than v_2, since he forfeits his bid even if he loses. Bidder 1 will take advantage of bidder 2's passivity by bidding less than v_2 himself, resulting in expected revenue strictly below v_2.[15] A cap on bids can attenuate bidder 1's ability to "preempt" bidder 2, which may increase bidding competition and raise the seller's expected revenue.

II. Equilibrium with a Cap on Bids

We now show the impact of an exogenous cap on bids. Let m denote the maximum allowable bid. We consider $m < v_2$ since a larger cap has no effect. There is a unique equilibrium in essentially all cases. That is, there is a unique pair of cumulative distribution functions for equilibrium bids. We focus on these cases first, followed by the nongeneric cases of $m = v_1/2$ and $m = v_2/2$. For small values of m, the equilibrium is in pure strategies. For larger values, it is in mixed strategies, although it still differs qualitatively from the equilibrium in the absence of a cap. (There are mass points at bids other than zero and there are gaps in the support of the equilibrium bids.) Let $F_i(z) \equiv \text{prob}(b_i \leq z)$ denote the cumulative distribution function for bidder i's bids in equilibrium.

Three lemmas provide necessary conditions for the distributions of equilibrium bids. The first lemma, which is stated without proof, shows that there cannot be mass points in the interval $(0, m)$. A sketch of the proof follows. If both bidders have mass points at a particular bid between zero and m, then each has an incentive to move the mass higher, since the (conditional) probability of winning would then jump up. (The same argument holds if both have mass points at zero.) If exactly one bidder has a mass point at the bid, then the other bidder will not place density immediately below that bid, since it would be preferable to move such density above the mass point. But it is profitable to move the mass lower if the other bidder has no density just below the mass point. Since there is an incentive to change bids in both cases, there cannot be mass points in $(0, m)$ in equilibrium. Proposition 1 of Hillman and Riley (1989) contains a proof for the case without a cap on bids (i.e., $m = \infty$).

LEMMA 1: *Neither bidder has a mass point at any bid $b \in (0, m)$. At most one bidder has a mass point at zero.*

An implication of Lemma 1 is that bidder i's expected payoff from a bid $b \in (0, m)$ is $v_i F_j(b) - b$, since there is zero probability that bidder j will also bid b. Another implication is that there cannot be a pure-strategy equilibrium here unless both bidders bid m. We now determine the lower limit of bids made in equilibrium. The proofs of the next two lemmas are in the Appendix.

LEMMA 2: *If $m \in (v_2/2, v_2)$, both bidders have an infimum bid of zero. If $m < v_2/2$, both have an infimum of m.*

Lemma 2 implies that the equilibrium bids are $b_1 = b_2 = m$ when $m < v_2/2$. We now show that there is a gap in the set of possible equilibrium bids when $m \in (v_2/2, v_2)$, and that both bidders have mass points at the cap.

LEMMA 3: *Suppose that $m \in (v_2/2, v_2)$. There exists a constant b' such that both bidders place nonzero density on every $b \in (0, b']$ and zero density on every $b \in (b', m)$. Both bidders have mass points at m.*

The lemmas provide necessary conditions for equilibrium distribution functions. We now find the unique pair that satisfies these conditions. Consider $m \in (v_2/2, v_2)$. We first determine the distribution functions that make the bidders indifferent among all bids in $(0, b'] \cup \{m\}$, as required by Lemma 3. We then find the equilibrium value of b'.

---

[14] An implication of Lemma 1, which follows, is that the equilibrium is in mixed strategies. Thus, bidder 2 has a nonzero probability of winning.

[15] Suppose that bidder 1 wins with (*ex ante*) probability $\pi < 1$. By the standard revealed preference argument, his expected payment, e, satisfies $\pi v_1 - e \geq v_1 - v_2$ since he could win with probability one by bidding v_2. Thus, $e \leq \pi v_2 - (1 - \pi)(v_1 - v_2) < \pi v_2$. At the same time, bidder 2 will not pay more than $(1 - \pi)v_2$ for a probability $(1 - \pi)$ of winning, so the seller's expected revenue is strictly below v_2. By contrast, if the seller could hold an oral auction, she could raise revenue of v_2. This last observation coincides with our result that changing the structure of the auction by imposing a cap may increase expected revenue.

Since bidder 1 must be indifferent among all bids in $(0, b'] \cup \{m\}$, each bid in that set must yield the same expected payoff. That is,

$$(1)\quad v_1F_2(b) - b = v_1[F_2(b') + (1 - F_2(b'))/2] - m,$$

for all $b \in (0, b']$. The left-hand side gives the expected payoff from bidding $b \in (0, b']$, while the right-hand side corresponds to a bid of m. (When bidder 1 bids m, there is probability $1 - F_2(b')$ that bidder 2 also bids m. The tie is broken in bidder 1's favor with probability $1/2$.) Similarly, a bid $b \in (0, b']$ yields bidder 2

$$(2)\quad v_2F_1(b) - b = v_2[F_1(b') + (1 - F_1(b'))/2] - m.$$

We now use (1) and (2) to show that bidder 2 has mass at zero. Straightforward algebra implies

$$(3)\quad v_1[1 - F_2(b)] = v_2[1 - F_1(b)],$$

for all $b \in (0, b']$.[16] Lemma 1 states that the bidders cannot both have mass points at zero, so either $F_1(0) = 0$ or $F_2(0) = 0$. Since $v_1 > v_2$, (3) implies $1 - F_2(0) < 1 - F_1(0)$, so $F_1(0) = 0$ and $F_2(0) = 1 - (v_2/v_1)$.[17]

The distribution functions can now be specified for bids above zero. Bidder 2's equilibrium expected payoff is zero, since a bid of zero yields $v_2F_1(0) = 0$, so (2) implies

$$(4)\quad v_2F_1(b) - b = 0,$$

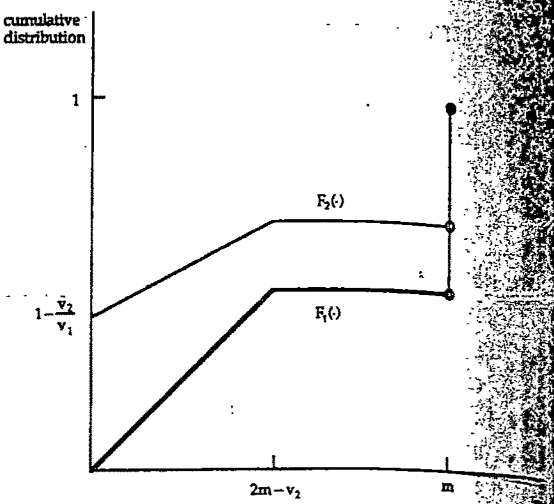


FIGURE 1. EQUILIBRIUM DISTRIBUTION FUNCTIONS

for all $b \in (0, b']$. Thus, bidder 1's distribution function satisfies $F_1(b) = b/v_2$ in [illegible] range. Lemma 3 implies $F_1(b) = F_1(b') =$ b'/v_2 for $b \in (b', m)$. Finally, $F_1(m) = 1$ [illegible] definition.

Bidder 1's equilibrium expected payoff [illegible] $v_1F_2(0) = v_1 - v_2 > 0$.[18] A bid $b \in (0, b']$ therefore gives bidder 1 an expected payoff [illegible]

$$(5)\quad v_1F_2(b) - b = v_1 - v_2,$$

so $F_2(b) = 1 - (v_2 - b)/v_1$ for $b \in [0, b']$. Finally, $F_2(b) = 1 - (v_2 - b')/v_1$ for $b \in (b', m)$. Equation (2) now implies that $b' =$ [illegible] $- v_2$. The equilibrium distribution functions are graphed in Figure 1.

All bids made by bidder 1 yield an expected payoff of $v_1 - v_2$, by construction, while all bids made by bidder 2 yield [illegible]. All other feasible bids are inferior to a bid [illegible] b' since a bid in (b', m) wins with the same probability as a bid of b', but is more costly. Thus, we have found the equilibrium bidding strategies.

---

[15] Rearranging (1) yields $m - b = v_1[(1 + F_2(b'))/2] - v_1F_2(b)$, so $m - b' = v_1[(1 - F_2(b'))/2]$. Adding these two equations yields $2m - b - b' = v_1[1 - F_2(b)]$. Repeating the exercise for bidder 2 yields $2m - b - b' = v_2[1 - F_1(b)]$.

[17] Since (3) holds only for $b > 0$, we use the fact that $\lim_{b\downarrow 0}F_i(b) = F_i(0)$.

[18] This holds since an infinitesimal bid gives [illegible] expected payoff of $\lim_{b\downarrow 0}v_1F_2(b) - b = v_1F_2(0)$ [illegible] v_2. Absent a cap, bidder 1 could guarantee a vi[illegible] bidding v_2, again yielding $v_1 - v_2$. The cap pre[illegible] from making that direct inference here.

We can now determine the seller's expected revenue. Note first that bidder 1's (*ex ante*) probability of winning the prize is

$$(6)\quad \int_0^{b'} \frac{1}{v_2} F_2(b)\,db + \left[1 - \frac{b'}{v_2}\right] \times \left[F_2(b') + \frac{1 - F_2(b')}{2}\right] = 1 - (v_2/2v_1).$$

The integral gives the probability that bidder 1 wins, conditional on bidding in $(0, b']$. [The density is $1/v_2$, and a bid b wins with probability $F_2(b)$.] The second term corresponds to the case in which he bids m. Bidder 1's expected payment is the difference between his gross and net expected payoffs:

$$(7)\quad v_1[1 - (v_2/2v_1)] - [v_1 - v_2] = v_2/2.$$

Using the same approach, bidder 2's expected payment can be expressed as $(v_2/v_1)(v_2/2)$. The seller's expected revenue is the sum:

$$(8)\qquad v_2(v_1 + v_2)/2v_1.$$

Now consider $m < v_2/2$. Lemma 2 implies that both bidders must bid m in any equilibrium. Bidding m is clearly equilibrium behavior. If $b_i = m$, then $b_j = m$ yields an expected payoff of $v_j/2 - m > 0$. A higher bid is not feasible while a lower bid loses with probability one, yielding a nonpositive expected payoff. In equilibrium, each bidder wins with probability $^1/_2$, and the seller's expected revenue is $2m$. We now summarize the results.

PROPOSITION 1: *If* $m \in (v_2/2, v_2)$, *bidder* 1 *wins with probability* $1 - (v_2/2v_1)$, *and the seller's expected revenue is* $v_2(v_1 + v_2)/2v_1$. *If* $m < v_2/2$, *bidder* 1 *wins with probability* $^1/_2$, *and the expected revenue is* $2m$.

A cap $m \in (v_2/2, v_2)$ leaves the seller's expected revenue the same as it is without a cap. If $m \in (v_2(v_1 + v_2)/4v_1, v_2/2)$, the expected revenue is $2m > v_2(v_1 + v_2)/2v_1$, so it exceeds revenue without a cap. The cap is small enough here to remove bidder 1's ability to preempt, but it is large enough that the increase in bidder 2's aggressiveness outweighs the decrease in bidder 1's. For lower m, the expected revenue is strictly lower with a cap than without. Expected revenue is graphed in Figure 2, as m varies.

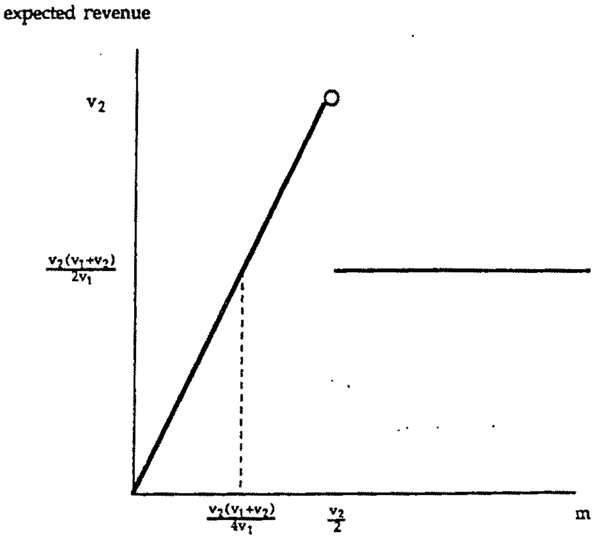


FIGURE 2. EXPECTED REVENUE

Total surplus is (weakly) lower with a cap than without. When $m > v_2/2$, bidder 1's probability of winning is the same as without the cap, so total surplus is unchanged. When $m < v_2/2$, however, bidder 1 wins with probability $^1/_2 < 1 - (v_2/2v_1)$, making total surplus strictly lower than without a cap.

We now consider the nongeneric cases of $m = v_1/2$ and $m = v_2/2$. If $m = v_2/2$, then there is an equilibrium in which $b_1 = b_2 = m$. There is also a continuum of equilibria in which bidder 2 places mass of v_2/v_1 or more on m and the remainder on zero, while bidder 1 always bids m.[19] If $m = v_1/2$, then there is an equilibrium of the form described in Lemma 3, but there is also an equilibrium in which bidder 1 randomizes between zero and m, while bidder 2 always bids m.[20]

---

[19] The proof of Lemma 2 showed that an infimum bid in $(0, m)$ was inconsistent with equilibrium, so only zero and m can be infimum bids here. The bidders can have different infimum bids only if one always bids m while the other randomizes between zero and m.

[20] The latter equilibrium requires $m = v_1/2 \le v_2$, or else bidder 2 is not optimizing. (The cap does not bind if $m =$

The qualitative results from these cases mirror the earlier results. Bidder 1's probability of winning is lower in these equilibria than in the case without a cap, so total surplus is again lower. The equilibria for $m = v_2/2$ again show that expected revenue can be higher with a cap than without.

III. Concluding Remarks

We have shown that an exogenous cap on bids in an all-pay auction (weakly) reduces the probability that the high-valuation bidder wins and increases the seller's expected revenue. When lobbying is seen as an all-pay auction, the results imply that limits on individual expenditures may increase total expenditures and lower total surplus. A cap on campaign contributions may therefore have the perverse effect of increasing aggregate contributions while lowering total surplus.

This paper also contributes to auction theory. We have characterized the equilibrium of the all-pay auction in the presence of an exogenous cap on bids. The results are applicable to a range of contests in which a limit is imposed on effort or expenditure, or in which contestants are constrained because of limited endowments. For instance, caps may increase total expenditures and lower total surplus in a war of attrition that would exhibit preemption in the absence of caps.[21] Limiting individual expenditures on research and development could increase total expenditures, and shorten the expected time to innovation. We conclude with a further discussion of robustness and implications of the results.

A. *Additional Bidders*

Suppose that there are $n > 2$ bidders, with valuations $v_1 > v_2 > \cdots > v_n$. In the absence of a cap, only the two bidders with the highest valuations make nonzero bids, so the expressions for expected revenue and total surplus are unchanged (Hillman and Riley, 19[illegible] Baye et al., 1993, 1996). Now suppose that there is a cap satisfying $v_k/k > m > v_{k+1}/(k + 1)$ for some $k < n$. There is an equilibrium with expected revenue of km. Bidding [illegible] gives a strictly positive expected payoff to bidders 1, 2, . . . , k. Any other feasible bid loses with probability one. At the same time, bidders $k + 1, k + 2, \ldots, n$ have no incentive to submit a nonzero bid.[22] Thus, expected revenue may again rise relative to the case without a cap. Total surplus is strictly lower than without a cap.

B. *Incomplete Information*

The results do not rely crucially on the assumption of complete information. Rather, the *ex ante* asymmetry in valuations generates the properties discussed above. While explicit handicapping of the high-valuation bidder may increase expected revenue in a first-price auction (Roger B. Myerson, 1981), a *symmetric* cap has that effect here. Thus symmetric limits on effort or expenditure can substitute for handicapping of favorites in tournaments and contests.

C. *Socially Wasteful Lobbying*

Lobbying expenditures take forms other than campaign contributions. For example, lobbyists spend money on public opinion surveys, and on print, radio, and television advertisements. They spend money encouraging citizens to participate in letter-writing campaigns. They also make in-kind contributions to politicians. Moreover, the value that a politician places on an in-kind contribution may be less than the expenditure involved. These observations suggest that not all lobbying expenditures accrue to politicians.

$v_1/2 > v_2$, so this case does not arise.) This equilibrium is unappealing in the sense that mass is put on a bid that is weakly dominated. Bidder 1 is indifferent between zero and m, but if bidder 2 might tremble and not bid m, then a bid of m would strictly dominate a bid of zero for bidder 1. The same point holds for the equilibria in which bidder 2 randomizes between zero and m.

[21] These preemption results are discussed in Drew Fudenberg and Jean Tirole (1991), for example.

[22] The identity of the active bidders is not uniquely determined if $v_{k+1}/k > m > v_{k+1}/(k + 1)$, since the[illegible] equilibrium in which bidder $k + 1$ bids m while [illegible] bids zero.

Our results do not depend critically on the assumption that bids accrue to the seller. Suppose that a fraction τ of each dollar spent on lobbying is effectively wasted, $0 < \tau < 1$. In the case of a political campaign, contributions enhance the politician's reelection chances, but the value to her of a contribution equal to b is only $(1 - \tau)b$. With two bidders and a cap $m < v_2/2$, expected revenue is $2(1 - \tau)m$, and total surplus is $(v_1 + v_2)/2 - 2\tau m$. There are now two ways that the cap can lower total surplus. In addition to increasing the probability that the prize goes to the low-valuation bidder, the cap may also increase the deadweight loss associated with the wasting of resources. In the region where the cap increases expected revenue, the amount of waste is higher with the cap than without, since waste increases with total expenditures.

D. *Divergence of Private and Social Valuations*

Whether the private and social valuations of the political prize are congruent in any particular context is important. Suppose that one bidder represents a corporation or a closely held business while the other represents a diffuse group, such as consumers. Free-riding among members of the latter group may lead to an understatement of their individual valuations and a consequent lowering of their probability of winning. Imposing a cap and increasing the group's probability of winning may therefore increase total surplus. Similar results hold if the groups differ in both valuations and costs of lobbying.

E. *Empirical Implications*

Aggregate spending on congressional races has doubled since passage of the FECA, which placed controls on campaign contributions. While consistent with the prediction of this paper, such evidence is not definitive since there have been other developments in the intervening years. It is also unclear how tightly the contribution limits bind since there are indirect ways to contribute. (For example, "soft money" can be funneled to political parties.) Thus, a careful empirical analysis is needed.

There are related studies the conclusions of which are consistent with the predictions of this paper. Several researchers have found that close electoral contests induce greater contributions (see James F. Herndon, 1982; Keith T. Poole et al., 1987, for example). To the extent that caps on campaign contributions make elections close, this finding is consistent with the prediction of our paper.

APPENDIX

PROOF OF LEMMA 2:

Let $b^* \equiv \inf\{z \mid F_1(z) > 0\}$ denote the infimum of bidder 1's bids. We first show that only zero or m can be infimum bids in equilibrium. Suppose instead that bidder 1's infimum bid is $b^* \in (0, m)$. If bidder 2 makes a bid in $(0, b^*)$, he loses with probability one. Since a bid of zero is better, bidder 2 must have zero density in $(0, b^*)$.[23] This means that bidder 1 could profitably move density in $(b^*, b^* + \varepsilon^*)$ arbitrarily close to zero. For $b \in (b^*, b^* + \varepsilon^*)$, the payment would drop by b. The probability of winning would drop by only $F_2(b) - F_2(0) = F_2(b) - F_2(b^*) < F_2(b^* + \varepsilon^*) - F_2(b^*)$, however. This last term is of order ε^*, by Lemma 1. It follows that moving the density raises bidder 1's expected payoff, for some $\varepsilon^* > 0$. Since a profitable deviation exists, an infimum bid of $b^* \in (0, m)$ cannot occur in equilibrium. The symmetric argument shows that bidder 2 cannot have an infimum in $(0, m)$ either, so only zero and m are possible infimum bids in equilibrium.

The remainder of the proof comprises two cases. Suppose first that $m \in (v_2/2, v_2)$. We employ a proof by contradiction to show that both bidders have an infimum of zero. Suppose instead that $b^* = m$, which implies that bidder 1 bids $b_1 = m$. Bidder 2 will bid zero or m, or he will randomize between the two, since a bid of zero strictly dominates any $b \in (0, m)$. Bidding zero is inconsistent with equilibrium since $b_1 = m$ is not optimal if $b_2 = 0$. Bidding m or randomizing between zero and m can only be optimal for bidder 2 if $v_2/2 - m \geq 0$, since a bid of m results in a tie. This restriction on m contradicts $m \in (v_2/2, v_2)$, so $b^* = m$ cannot occur in equilibrium here. An analogous argument shows that bidder 2 cannot have an infimum bid of m, so the common infimum is zero.

---

[23] This condition on the density function holds for *almost every* bid in the interval $(0, b^*)$. That is, it can be violated on a set of measure zero without changing the result. In the interest of brevity, we leave this qualification understood throughout the paper.

Now suppose that $m < v_2/2$. We employ a proof by contradiction to show that both bidders have an infimum of m. Bidding m guarantees at least a tie, so bidder i must receive an expected payoff of at least $v_i/2 - m > 0$. Suppose that bidder i has an infimum bid of zero. Since his infimum is zero, a bid near zero must be as good as a bid of m for bidder i. But if bidder j does not have mass at zero, then bidder i receives less than $v_i/2 - m$ if he bids near zero. (The probability that bidder i wins would be arbitrarily small for bids that are arbitrarily close to zero.) Bidder j must therefore have mass at zero. Since bidder j's infimum is also zero, the same argument implies that bidder i must have mass at zero. The bidders cannot both have mass at zero, by Lemma 1, so the infimum must equal m for both bidders.

PROOF OF LEMMA 3:

We first show that both bidders have mass points at m. Lemma 2 shows that the common infimum is zero, while Lemma 1 shows that at least one bidder has no mass at zero. Suppose, in particular, that bidder i does not have mass at zero. Then, if bidder j bids arbitrarily close to zero, his expected payoff is strictly below $v_j - m > 0$. Since his infimum is zero, a bid near zero must be as good as a bid of m for bidder j. But if bidder i does not have mass at m, then a bid of m would yield bidder j an expected payoff of $v_j - m$. Bidder i must therefore have mass at m. We conclude that at least one bidder has mass at m, since at least one bidder has no mass at zero.

Suppose, in particular, that bidder i has mass $\alpha > 0$ at m. If bidder j has nonzero density in $(m - \varepsilon', m)$, he could profitably move it to m, for some $\varepsilon' > 0$. For $b \in (m - \varepsilon', m)$, the payment would rise by only $m - b < \varepsilon'$, but the probability of winning would rise by at least $\alpha/2$, since bidder j would now tie if bidder i bids m. Since moving the density up raises bidder j's exp
payoff, bidder j must have zero dens
$(m - \varepsilon', m)$. If bidder j has no mass
then bidder i could profitably take mass
m and move it lower. We conclude that
bidders have mass points at m.

The presence of mass points at m for
bidders implies that both bidders have
density in $(m - \varepsilon'', m)$, for some $\varepsilon'' > 0$.
demonstrates the existence of $b^* \in [0,$
such that both bidders have zero density
(b^*, m). Let b' denote the smallest $b^* \in$
$m]$ such that both bidders place zero
on every bid in (b^*, m).[24] We now show
both bidders place nonzero density on
$b \in (0, b']$. By the argument used in the
of Lemma 2, if bidder i has zero density
interval $(s, t) \subset (0, b']$, then so must
j. But if both bidders have zero density
$t)$, then either bidder could profitably
density from $(t, t + \varepsilon^*)$ down to s, for
$\varepsilon^* > 0$. Thus, both bidders must have no
density on every $b \in (0, b']$.

REFERENCES

Alexander, Herbert E. and Bauer, Monica. *nancing the 1988 election*. Boulder,
Westview Press, 1991.

Austen-Smith, David. "Campaign Contr
tions and Access." *American Political*
ence Review, September 1995, *89*(3),
566–81.

Austen-Smith, David and Wright, John R. "
petitive Lobbying for a Legislator's Vo
Social Choice and Welfare, July
9(3), pp. 229–57.

Baye, Michael R.; Kovenock, Dan and de
Casper G. "Rigging the Lobbying Pro
An Application of All-Pay Auction
American Economic Review, March
83(1), pp. 289–94.

______. "The All-Pay Auction with Com
Information." *Economic Theory*, Au
1996, *8*(2), pp. 291–305.

---

[24] In fact, $b' \in (0, m)$. The preceding paragraph
that $b' \neq m$. If $b' = 0$, then both bidders must have
points at zero, by Lemma 2. This contradicts
so $b' \neq 0$.

Congressional Quarterly. "House Votes to Toughen Gift Restrictions." November 18, 1995a, *53*(45), pp. 3516–19.

———. "Bill Would Open Windows on Lobbying Efforts." December 2, 1995b, *53*(47), pp. 3631–33.

———. "Hoekstra Proposes Legislation to Make Overhaul GOP Issue." February 3, 1996, *54*(5), p. 281.

Fudenberg, Drew and Tirole, Jean. *Game theory*. Cambridge, MA: MIT Press, 1991.

Hall, Richard L. and Wayman, Frank W. "Buying Time: Moneyed Interests and the Mobilization of Bias in Congressional Committees." *American Political Science Review*, September 1990, *84*(3), pp. 797–820.

Helsley, Robert W. and O'Sullivan, Arthur. "Altruistic Voting and Campaign Contributions." *Journal of Public Economics*, September 1994, *55*(1), pp. 107–19.

Herndon, James F. "Access, Record, and Competition as Influences on Interest Group Contributions to Congressional Campaigns." *Journal of Politics*, November 1982, *44*(4), pp. 996–1019.

Hillman, Arye L. and Riley, John G. "Politically Contestable Rents and Transfers." *Economics and Politics*, Spring 1989, *1*(1), pp. 17–39.

Levitt, Steven D. "Congressional Campaign Finance Reform." *Journal of Economic Perspectives*, Winter 1995, *9*(1), pp. 183–93.

Myerson, Roger B. "Optimal Auction Design." *Mathematics of Operations Research*, February 1981, *6*(1), pp. 58–73.

Poole, Keith T.; Romer, Thomas and Rosenthal, Howard. "The Revealed Preferences of Political Action Committees." *American Economic Review*, May 1987 (*Papers and Proceedings*), *77*(2), pp. 298–302.

Romer, Thomas and Snyder, James M., Jr. "An Empirical Investigation of the Dynamics of PAC Contributions." *American Journal of Political Science*, August 1994, *38*(3), pp. 745–69.

Tullock, Gordon. "Efficient Rent Seeking," in James M. Buchanan, Robert D. Tollison, and Gordon Tullock, eds., *Toward a theory of the rent seeking society*. College Station, TX: Texas A&M University Press, 1980, pp. 269–82.

Wright, John R. "Contributions, Lobbying and Committee Voting in the U.S. House of Representatives." *American Political Science Review*, June 1990, *84*(2), pp. 417–38.

Is There a Free-Rider Problem in Lobbying? Endogenous Tariffs, Trigger Strategies, and the Number of Firms

By PAUL PECORINO*

> On industry structure, it is perhaps disappointing that the empirical literature is not more clearcut on the political advantages of high concentration, in view of the strong presumption that free-rider effects should be important in lobbying. Here too, however, theoretical contributions are scant.
>
> —(Dani Rodrik, 1995 p. 1483)

There are two reasons we expect more-concentrated industries to have advantages over less-concentrated industries in playing the lobbying game.[1] One is that in a noncooperative setting, as the number firms rises, we expect the free-rider problem to become worse (this is the finding of Rodrik, 1986). The second is that we expect that the ability to maintain a cooperative outcome deteriorates as the number of firms in the industry rises. [The classic reference is Mancur Olson (1965); also, see Russell Hardin (1982) and Todd Sandler (1992) among others.]

In contrast with the second intuition, this paper shows that there is no general presumption that maintaining cooperation becomes more difficult as the number of firms in the industry rises. The ability of an industry to overcome the free-rider problem is analyzed in a repeated tariff lobbying game where cooperation may be maintained through the use of a simple trigger strategy.[2] The difficulty of maintaining cooperation is measured by the critical value of the discount parameter. All values above this critical value are consistent with cooperation, and all values below the critical value imply that cheating will prevent maintenance of the cooperative outcome. There is no presumption that this critical value rises with the number of firms. A separate, but related, question involves asymptotic results. Standard intuition suggests that as industry structure approaches perfect competition, it becomes impossible to maintain cooperation. In the context of this paper, this intuition corresponds to the critical value of the discount factor converging to 1 as the number of firms increases to infinity. In the general case it is shown that the critical value of the discount parameter converges to a value less than 1 in the limit as the number of firms becomes arbitrarily large. If the actual discount parameter lies above this value, then cooperation may be maintained even with an infinite number of firms.

Though the framework here is quite simple, it provides results which contrast sharply with our standard intuition about how industry structure affects the ability to maintain cooperation.[3] There are two ways to interpret the results of this paper. One is to assert that standard intuition is wrong, and that there is, in fact, no presumption that maintaining cooperation becomes more difficult as the num-

---

* Department of Economics, Finance and Legal Studies, University of Alabama, Box 870224, Tuscaloosa, AL 35487. I would like to thank Dan Arce, John Conlon, Amy Farmer, Costas Syropoulos, Akram Temimi, and three anonymous referees for helpful comments and discussions regarding the paper. This research was supported by a grant from the University of Alabama's College of Commerce and Business Administration.

[1] In this paper, I will consider industries where firms are identical. Thus I will use firm concentration and the number of firms in the industry interchangeably.

[2] Trigger strategies for infinitely repeated games were first analyzed by James W. Friedman (1971). See John C. Conlon (1996) for a recent analysis of cooperation [illegible] finitely repeated setting.

[3] For example, in a standard infinitely repeated Bertrand oligopoly game with homogeneous goods and constant marginal costs, the critical value of the discount parameter is $(n - 1)/n$, where n is the number of firms in the industry. See Carl Shapiro (1989 p. 370).

of firms in an industry rises. As pointed out by the opening quote from Rodrik (1995), the empirical literature to date has not decisively established the advantages of concentration for receiving protection.[4] Another approach is to maintain that our intuition is correct, but that we need to think more deeply about the true nature of the monitoring technology. This might entail amending the simple trigger strategy model presented here, or outright rejection of the trigger strategy model in favor of some other enforcement mechanism, such as Olson's (1965) by-product theory.[5]

The trigger strategy framework might be extended to consider imperfect information.[6] In this paper, it is assumed that firms view the previous period's aggregate contribution without error. This does not seem highly problematic. In an industry with a large number of firms, a greater problem is likely to be that the number and size distribution of firms may not be known with certainty. Thus individual firms will have difficulty knowing what the cooperative level of contributions *should* be. Getting the cooperative level of contributions exactly right is not crucial, but establishing some level imposes a severe coordination problem, and if firms are heterogeneous, there will be problems establishing the cooperative share for each firm.

These caveats suggest that further research is needed on the question of how the number of firms in an industry is related to its ability to overcome the free-rider problem. What this paper demonstrates, however, is that there is nothing in the underlying structure of payouts which guarantees that cooperation must break down as the number of firms grows large.

Raymond Riezman and John D. Wilson (1995) analyze tariff lobbying in a repeated game setting, and suggest analyzing the effect of increases in the number of firms on the ability to maintain cooperation. The tariff lobbying problem is related to the more general problem of public goods provision. See the work of Sandler (1992) and Richard Cornes and Sandler (1996), among others. Also, see John McMillan (1979) for a treatment of public goods provision in a repeated game setting. The issues addressed in this paper have previously been addressed in the literature on oligopoly. Val E. Lambson (1984, 1987) lays out the rather stringent conditions under which collusion may be maintained in a repeated game oligopoly setting as the number of firms rises to infinity.[7]

The cooperative and noncooperative equilibria of the model are set out in Section I. In Section II, trigger strategies are analyzed with an emphasis on the effect of an increase in the number of firms on the critical value of the discount parameter, including the case where the number of firms increases without bound. Section III concludes.

I. Cooperative and Noncooperative Equilibria[8]

Firms in this model are defined by ownership of sector-specific capital.[9] In order to better focus on the problem at hand, the productive structure is made as simple as possible. In particular, industry output is linear in sector-specific capital, which is the only input. Further, units are chosen so that 1 unit of sector-specific capital produces 1 unit of output (see the Appendix for an analysis with a more general production technology). Industry output is homogeneous, and the good is traded internationally with a world price of 1. Thus, the domestic price is given by

$$P = 1 + t, \tag{1}$$

where the height of the tariff is denoted t.

---

[4] This indeterminacy is borne out in the Jan Potters and Randolph Sloof (1996) survey of the empirical literature on interest groups.

[5] In addition to Olson, see the discussions of this theory in [illegible] (1982) and Sandler (1992).

[6] See Edward J. Green (1980 Sec. 5), Robert H. Porter ([illegible]), Green and Porter (1984), and Dilip Abreu et al.

[7] See Shapiro (1989) for an extensive discussion of oligopoly in a repeated play setting.

[8] The noncooperative equilibrium for the static tariff game is considered in Rodrik (1986). Rodrik's model is a simplified version of Ronald Findlay and Stanislaw Wellisz (1982).

[9] There is no free entry or exit in the model, and therefore no zero profit condition is imposed. Riezman and Wilson (1995) discuss trigger strategies in the case of free entry and in particular consider the possibility of entry itself triggering a return to the noncooperative outcome.

The height of the tariff is determined by the sum of lobbying contributions s_i of each of the n firms in the industry. Thus, the tariff is given by[10]

$$(2) \qquad t = t(S), \qquad \text{where } S = \sum_{i=1}^{n} s_i .$$

It is assumed that $t(0) = 0$, $t'(S) > 0$, $t''(S) < 0$, and that $t'(S)$ is finite for all $S > 0$ (possibly approaching ∞ as S approaches 0).

Increases in the number of firms are considered holding industry capital constant. All firms are assumed to be identical, and only symmetric equilibria are considered throughout the analysis. Total industry capital is normalized to 1 making each firm of size $1/n$. Profits of firm i are

$$(3) \qquad \pi_i = (1/n)(1 + t(S)) - s_i .$$

The first-order condition for firm i implies

$$(4) \qquad t'(S_{-i} + s_i) \leq n,$$

$$s_i \geq 0 \textit{ with complementary slackness,}$$

where S_{-i} are lobbying contributions by all firms other than firm i. From (4) we can solve for S^N, the noncooperative level of tariff lobbying, which implies the noncooperative level of the tariff, t^N. A symmetric equilibrium is imposed in which $s_i^N = S^N/n$. This allocation of the lobbying burden is not unique, but the value of S^N which solves (4) is unique because the marginal cost of lobbying is constant, and the marginal benefit is monotonically decreasing in S^N.

In the cooperative equilibrium, firms jointly maximize

$$(5) \qquad \Pi = \sum_{i=1}^{n} \pi_i = 1 + t(S) - S,$$

where Π denotes industry-level profits. Th first-order condition to the cooperative pr lem implies

$$(6) \qquad t'(S) = 1.$$

From (6), we can solve for the unique va of S^C which, in turn, implies t^C.[11] With sy metry, $s_i^C = S^C/n$. The comparative stati with respect to the number of firms n are give by (7) and (8), where the N superscript ind cates the noncooperative equilibrium, and C superscript the cooperative equilibrium.

$$(7a) \qquad \frac{dS^N}{dn} = \frac{1}{t''(S)} < 0;$$

$$(7b) \qquad \frac{dt^N}{dn} = \frac{t'(S)}{t''(S)} < 0;$$

$$(8a) \qquad \frac{dS^C}{dn} = 0;$$

$$(8b) \qquad \frac{dt^C}{dn} = 0.$$

Equation (7) reflects the worsening of free-rider problem as n rises. Equation (8) re sults from the fact that industry size is he constant throughout the analysis. As a res the optimal (from the industry's perspective lobbying effort is independent of n. By co trast, in most analyses of public goods pro sion, when the number of individuals rises, optimal level of provision will also rise (Cornes and Sandler, 1996 pp. 161–63).

II. Trigger Strategies and the Number of Fi

A. *General Conditions*

Consider the possibility of cooperation in infinitely repeated game, where cooperatio supported through the use of a trigger strateg In such an equilibrium, each firm makes cooperative level of lobbying contribution, S^C/n, in the current period if all firms cooperated in the previous period. If any firm defected in the previous period, then all firms revert to the single-shot Nash equilibrium forever. Assume that all firms observe, without error, the aggregate level of contributions made in the previous period. As a result, defections are always detected. Let an individual firm's profit be denoted π^C under the cooperative equilibrium, and π^N under the noncooperative (i.e., single-shot Nash) equilibrium, respectively. The profit earned by a single firm which defects from a cooperative equilibrium is denoted π^D.

---

[10] Note that if more firms represent more workers, then it might be reasonable to assume some dependence of the tariff function on n. In our case, however, total industry size is held constant so that when there are a greater number of firms, each is smaller. Thus it is reasonable to assume that the tariff function is independent of n.

[11] An interior equilibrium is assumed under co tion. While it is conceivable that the optimal coop level of lobbying is zero, analyzing the effect of in of the number of firms on the ability to maintain co ation is not interesting in this case.

If a firm defects from the cooperative equilibrium, it will earn π^D for the current period, and π^N in all future periods. If this payoff is greater than the payoff from continued cooperation, then the firm will defect, and the cooperative equilibrium cannot be maintained. Thus, a necessary condition for maintaining a cooperative outcome under a trigger strategy with infinite Nash reversion is

$$\pi^D + \sum_{t=1}^{\infty} \delta^t \pi^N \leq \sum_{t=0}^{\infty} \delta^t \pi^C, \tag{9}$$

where δ denotes the discount factor. Define δ^* to be the critical value of δ such that (9) holds as a strict equality. For all $\delta \geq \delta^*$, the cooperative equilibrium can be supported, while for $\delta < \delta^*$, cooperation cannot be supported by any simple trigger strategy.[12] Evaluate the summations in (9) and solve for δ^* to get

$$\delta^* = \frac{\pi^D - \pi^C}{\pi^D - \pi^N}. \tag{10}$$

Note that since, $\pi^D \geq \pi^C \geq \pi^N$, $\delta^* \leq 1$.

The effect of an increase in the number of firms in the industry on the difficulty of maintaining cooperation will be measured through δ^*.[13] If an increase in n increases δ^*, then it is said to make cooperation more difficult. A related, but distinct, issue concerns the limit of δ^* as n becomes arbitrarily large. If δ^* approaches 1, then (absent the unrealistic case where there is no discounting of future profits) cooperation becomes impossible as n grows large. If δ^* approaches a fraction less than 1 in the limit as n grows large, then there are values of the discount parameter such that cooperation can be maintained with an infinite number of firms in the industry.

B. *Cooperation and the Number of Firms*

Expressions for π^C and π^N may be obtained from (3), using a superscript to indicate the appropriate equilibrium, while noting that $s_i^C = S^C/n$ and $s_i^N = S^N/n$. When a firm defects from the cooperative outcome, there are two cases. In one case, the firm cuts its contributions below the cooperative level to some positive level s^D dictated by equation (4). In the other case, the firm cuts its contribution to zero. This will be true if $t'(S_{-i}) < n$, where S_{-i} represents lobbying contributions by all firms other than firm i. We will analyze the latter case since it will hold except possibly when n is very small (i.e., when an individual firm's contributions make up a large percentage of the total). The assumption that $s^D = 0$ has no effect on the results, and the case where $s^D > 0$ is analyzed in Pecorino (1996).[14] Thus we have

$$\pi^C = (1/n)(1 + t^C - S^C); \tag{11a}$$

$$\pi^N = (1/n)(1 + t^N - S^N); \tag{11b}$$

$$\pi^D = (1/n)(1 + t^D). \tag{11c}$$

---

[12] Even when δ is sufficiently high, the cooperative equilibrium is clearly not unique (e.g., noncooperation is always an equilibrium).

[13] The focus in this paper is the effect of n on δ^*. As an alternative approach, suppose $\delta < \delta^*$, so that full cooperation cannot be maintained. Letting δ replace δ^*, we use (13) to implicitly define S^M, the maximum sustainable level of lobbying contributions, as a function of δ and n. Comparative statics can be used to compute dS^M/dn which shows how the maximum level of sustainable cooperation is affected by the number of firms in the industry.

[14] It is straightforward to show that $s^D \rightarrow 0$ as n grows large. As a result, assuming $s^D = 0$ has no effect on the asymptotic properties of the model. Pecorino (1996) is the working paper version of this paper.

Since $s_i^C = S^C/n$, it follows that $S_{-i} = ((n-1)/n)S^C$ and

$$(12) \qquad t^D(S_{-i}) = t\left(\frac{(n-1)}{n}S^C\right).$$

Substitute (11) into (10) to get

$$(13) \quad \delta^* = \frac{t^D - (t^C - S^C)}{t^D - (t^N - S^N)}$$

$$= \frac{t\left(\frac{(n-1)}{n}S^C\right) - (t(S^C) - S^C)}{t\left(\frac{(n-1)}{n}S^C\right) - (t(S^N) - S^N)}.$$

In general it is not possible to say what will happen to δ^* as n rises. The numerator will tend to rise with n. From (8), both S^C and t^C are invariant to n, but from (12), we know that t^D is increasing in n (t^D approaches t^C from below as n rises). The denominator also rises with n. As just discussed, t^D is increasing in n, and it is straightforward to show that $t^N - S^N$ (the net benefit to the industry in the noncooperative equilibrium) is decreasing in n.[15]

The sign of $d\delta^*/dn$ is the same as the sign of

$$\left(\frac{dt^D}{dn}\right)[(t^C - S^C) - (t^N - S^N)]$$

$$+ \left(\frac{dt^N}{dn} - \frac{dS^N}{dn}\right)[t^D - (t^C - S^C)].$$

As discussed in the paragraph above, dt^D/dn is positive, and $(dt^N/dn) - (dS^N/dn)$ is negative. The first term in brackets is positive because the net benefits to the industry are higher under cooperation than under noncooperation. The second term in brackets is positive because defection gives a higher single-pe payout than cooperation. Cooperation comes more difficult as n rises only if the term dominates the second. The first term flects a greater desirability of defection an increase in n because the tariff under fection is growing closer to the tariff cooperation. The second term reflects a grea penalty associated with defection because noncooperative equilibrium becomes less tractive as n rises. In general, there is no to know which of these terms will domin Thus, it is important to note that for the gener tariff function, $t(S)$, we do not have a the retical presumption that cooperation becom more difficult as the number of firms ri This is summarized as Result 1.

RESULT 1: *For a general tariff function the form* $t = t(S)$, *where* $t' > 0$, *and* $t'' <$ *the sign of* $d\delta^*/dn$ *is indeterminate. Ther fore, there is no presumption that* δ^* *(our mea sure of the difficulty of maintaining cooperation) rises with the number of firms.*

Result 1 has been derived assuming that S represents full cooperation among the firms. is possible, using an analog to (13), to de a "critical" discount parameter for any co stant level of cooperative contributions S su that $S^N < S \leq S^C$, where S^C is full coopera It is straightforward to show that the analog Result 1 applies to any constant level of operation within this range.

It has been assumed that there are no fi set-up costs for lobbying each period. In presence of such costs, there will be s modification to Result 1. Suppose fixed co K must be incurred each period before bying contributions have any impact on tariff at the margin. Since the net benefits f lobbying in the noncooperative case go to n rises to ∞ (see the discussion in subsec C below), there will be some finite n, ca n^K, above which the net benefits of lobb fall below the fixed costs K. As a result level of noncooperative lobbying will rev a corner solution with zero contributions a zero tariff. For $n \geq n^K$, maintaining co eration will become more difficult as n As noted above, there are two effects on n rises. One is that the noncooperative eq

---

[15] Consider the special case where $t^C = (S^C)^\alpha$. Numerical simulations reveal that for $\alpha < 0.54$, δ^* is monotonically decreasing in n, for $\alpha > 0.67$ δ^* is monotonically increasing in n, while for α between 0.54 and 0.67, the relationship between δ^* and n is not monotonic.

[illegible] becomes less attractive, decreasing both the incentive to cheat, and δ^*. Once a corner solution is reached at n^K, the noncooperative equilibrium is as bad as it will get, and this effect disappears. The remaining effect is that the tariff under defection rises towards the tariff under cooperation, which causes the incentive to cheat, and δ^* to both rise. For a fuller treatment of this issue, see Pecorino (1996).

C. *Asymptotic Results*

Another interesting question is what happens to δ^* as n becomes arbitrarily large. If δ^* approaches 1, and firms do any discounting at all, it is guaranteed that cooperation will break down as n grows large. If δ^* approaches a fraction less than 1, then there may be plausible values of the discount parameter δ which are consistent with cooperation being maintained with an infinite number of firms. In the model, there is no presumption that cooperation breaks down as the number of firms in the industry rises without bound.

To see this, consider the expression for δ^* in equation (13). We would like to take the limit of this expression as $n \rightarrow \infty$. In doing so, consider some of the individual terms in the fraction in (13). From (12) note that $t^D \rightarrow t^C$ as $n \rightarrow \infty$. Also note that $S^N \rightarrow 0$ as $n \rightarrow \infty$. This follows from the first-order condition in (4), and the fact that $t'(S)$ is finite for $S > 0$ [for t' infinite, (4) cannot hold as an equality for any positive level of S]. Thus as n grows without bound, the level of noncooperative lobbying must approach zero.[16] Since $t(0) = 0$, this implies that $t^N - S^N \rightarrow 0$ as $n \rightarrow \infty$. Bearing all this in mind, we have,

$$\lim_{n \rightarrow \infty} \delta^* = \frac{S^C}{t^C} \leq 1. \tag{14}$$

Recall that S^C and t^C are invariant to n. This expression cannot be greater than 1 because this would imply industry losses from lobbying under the cooperative equilibrium. The industry can always ensure zero losses from lobbying by setting $S = 0$. The expression in (14) is less than 1 when there is an interior solution which yields the industry positive net benefits (i.e., $t^C - S^C > 0$) in the cooperative equilibrium (if this expression is not less than 1, the question of what happens when n rises is entirely moot).

In the limit, $(1/n)S^C$ is the temptation to cheat faced by a firm, while $(1/n)t^C$ is the per-period future benefit from maintaining cooperation. The critical value of δ approaches the ratio of these two terms in the limit as n grows large. Since this ratio is less than 1 when there is an interior solution, there is no presumption that cooperation must break down as the number of firms in the industry grows arbitrarily large. This strongly contrasts with standard intuition, and is stated as Result 2.

RESULT 2: *If the cooperative equilibrium yields positive net benefits to the industry, then in the limit as the number of firms in the industry grows without bound,* δ^* *converges to* $S^C/t^C < 1$. *If* $\delta > S^C/t^C$, *the cooperative outcome may be maintained with an infinite number of firms in the industry (i.e., under conditions of perfect competition).*

Result 2 implies that there are admissible discount parameters which allow full cooperation to be maintained in the limit. It is straightforward to show that an analog to Result 2 holds for a constant cooperative level of contributions S in the range $S^N < S \leq S^C$, where S^C represents full cooperation.

Result 2 is unaffected by the consideration of fixed set-up costs of lobbying. As discussed in subsection B above, with fixed set-up costs of lobbying, there is a finite n above which both contributions S^N and the tariff t^N in the noncooperative equilibrium go to zero. However, even in the absence of fixed set-up costs of lobbying, t^N and S^N approach zero asymptotically.

III. Conclusion

The attractiveness of the framework adopted here is its simplicity; a trigger strategy is a very simple monitoring technology. However, this framework gives results which are strongly at odds with our intuition about how increases in the number of firms in an industry affect the probability of overcoming the free-rider problem inherent in tariff lobbying. In particular, there is no presumption in the model that overcoming the free-rider problem becomes more difficult as the number of firms in the industry rises. Also, there is no presumption that cooperation must break down as the number of firms increases without bound. The crucial feature of the model is the way the single-period payouts π^D, π^C, and π^N change as n rises, and these changes are independent of the length of the punishment period. Thus, the results of the model should not depend crucially on assumption of infinite punishment periods.[17] Note that if punishments more sophisticated than reversion to the single-shot Nash equilibrium are considered, as in Abreu (1988), the results of this paper should continue to hold. If a simple strategy is able to maintain cooperation as the number of firms grows large, then an "optimal penal code" as characterized by Abreu should do at least as well.[18] The results may not be robust to the introduction of certain types of informational imperfections, and this should be an area of future research.

---

[16] The same result is found in Rodrik (1986).

The standard intuition about industry structure and the ability to maintain cooperation in lobbying is not supported by a simple model in which cooperation is maintained through the use of a trigger strategy. Either our intuition is wrong, or if it is correct, we must appeal to more complicated models of the lobbying process which might explicitly include imperfect information, or transaction costs among other features.

APPENDIX

In the main body of the text, firms sumed to have a fixed production where output is simply equal to the size firm's capital stock. In this Appendix, consider firms defined by ownership of a amount of sector-specific capital, but wit cess to a constant-returns-to-scale produ technology using labor as a variable With firm capital fixed, the production tion will exhibit diminishing returns in The focus in this Appendix is on asym results.

Consider firms with constant-returns scale production functions $f(l_i, 1/n)$, whe is labor hired by firm i at the wage rate w, $1/n$ is the capital owned by the firm. Profi firm i are $\pi^i = (1 + t)f(l_i, 1/n) - wl_i$ After the tariff has been set, firm i hires until

$$(\text{A1}) \qquad (1 + t)f_{l_i}\left(l_i, \frac{1}{n}\right) = w,$$

where f_{l_i} is the partial derivative of the duction function with respect to the input.

Note that because of the constant-return scale production function, if domestic held constant, industry output $F = \Sigma$ invariant to the number of firms n. Where venient, the industry's aggregate produ function will be denoted $F(L, 1)$, wher reflects total industrywide capital, and gregate industry employment. In the operative equilibrium firm i lobbies until

$$(\text{A2}) \qquad t'(S_{-i} + s_i)f\left(l_i, \frac{1}{n}\right) \leq 1,$$

$$s_i \geq 0 \textit{ with complementary slackness}$$

As in the main body of the paper, in the cooperative equilibrium, the tariff appro zero as n grows large ($f^i \to 0$ as $1/n$ yielding the free-trade equilibrium. Due constant-returns-to-scale technology, to dustry employment, output, and profits free trade are invariant to n. If we deno dustry employment, output, and profits free trade as L^F, F^F and Π^F, then asym

---

[17] For punishment periods which are less than infinite, δ^* will be higher for any given n; however the change in δ^* as n rises will be governed by the changes in π^D, π^C, and π^N.

[18] The most severe punishment which can be imposed on a defector is for all other firms to contribute zero in all periods after the initial defection. Because the noncooperative level of the tariff is the tariff a single firm would find it individually rational to lobby for, the punished firm would rationally contribute S^N in all punishment periods providing a tariff of t^N. Under such a punishment scheme, Results 1 and 2 of the paper will continue to hold. Note that this punishment would require observing the actions of all firms individually rather than just the aggregate contribution.

Similarly, each firm receives $\pi^N = (1/n)\Pi^F$ in a free-trade equilibrium. It will prove convenient to write this as

$$\text{(A3)} \quad \lim_{n\to\infty} n\pi^N = \Pi^F = F^F - wL^F.$$

Under cooperation, firms lobby to maximize industry profits:

$$\Pi = \sum_{i=1}^{n} \pi^i = (1 + t)F(L, 1) - wL - S.$$

The first-order condition to this problem implies

$$\text{(A4)} \qquad t'F(L, 1) = 1.$$

From (A1) and (A4) it is possible to solve for t^C and L^C, and these, in turn, imply values of S^C, Π^C, and F^C. Each of these industry-level variables is invariant to n. This follows since output is independent of n holding the tariff constant. This means that the optimal level of the tariff [see (A4)] is independent of n. Each firm receives $\pi^C = \Pi^C/n$, which may be rewritten as

$$\text{(A5)} \quad n\pi^C = \Pi^C = (1 + t^C)F^C - wL^C - S^C.$$

Note that output and employment of each firm are given by $f^C = F^C/n$ and $l^C = L^C/n$.

Let the profits of a firm which defects be denoted (in the case where $s^D = 0$)

$$\text{(A6)} \quad \pi^D = (1 + t^D)f(l_i^D, 1/n) - wl_i^D.$$

As in the main text, $\lim_{n\to\infty} t^D = t^C$ because an individual firm's withheld contribution grows smaller and smaller as n rises. Since the tariff under defection approaches the tariff under cooperation, it follows that output and employment of the defecting firm approach their levels under cooperation, i.e., $\lim_{n\to\infty} l^D = l^C = L^C/n$ and $\lim_{n\to\infty} f^D = f^C = F^C/n$. Thus,

$$\text{(A7)} \quad \lim_{n\to\infty} n\pi^D = (1 + t^C)nf^C - wnl^C$$

$$= (1 + t^C)F^C - wL^C$$

$$= \Pi^C + S^C.$$

From (10), we can write $\delta^* = (n\pi^D - n\pi^C)/(n\pi^D - n\pi^N)$. Use this plus (A3), (A5), and (A7) to obtain

$$\text{(A8)} \quad \lim_{n\to\infty} \delta^* = \frac{S^C}{S^C + \Pi^C - \Pi^F} < 1.$$

Recall that $\Pi^C - \Pi^F > 0$ since industry profits under cooperation are greater than industry profits under free trade. Result 2 still holds when we allow firms access to constant-returns-to-scale, variable input production technologies.

REFERENCES

Abreu, Dilip. "On the Theory of Infinitely Repeated Games with Discounting." *Econometrica*, March 1988, *56*(2), pp. 383–96.

Abreu, Dilip; Pearce, David and Stacchetti, Ennio. "Optimal Cartel Equilibria with Imperfect Monitoring." *Journal of Economic Theory*, June 1986, *39*(1), pp. 251–69.

Conlon, John R. "Cooperation for Pennies: A Note on ε-Equilibria." *Journal of Economic Theory*, August 1996, *70*(2), 489–500.

Cornes, Richard and Sandler, Todd. *The theory of externalities, public goods, and club goods*, 2nd Ed. Cambridge: Cambridge University Press, 1996.

Findlay, Ronald and Wellisz, Stanislaw. "Endogenous Tariffs, the Political Economy of Trade Restrictions, and Welfare," in Jagdish N. Bhagwati, ed., *Import competition and response*. Chicago: University of Chicago Press, 1982, pp. 223–38.

Friedman, James W. "A Non-cooperative Equilibrium for Supergames." *Review of Economic Studies*, January 1971, *38*(1), pp. 1–12.

Green, Edward J. "Noncooperative Price Taking in Large Dynamic Markets." *Journal of Economic Theory*, April 1980, *22*(2), pp. 155–82.

Green, Edward J. and Porter, Robert H. "Noncooperative Collusion Under Imperfect Price Information." *Econometrica*, January 1984, *52*(1), pp. 87–100.

Hardin, Russell. *Collective action*. Baltimore, MD: Johns Hopkins University Press, 1982.

Lambson, Val E. "Self-Enforcing Collusion in Large Dynamic Markets." *Journal of Economic Theory*, December 1984, *34*(2), pp. 282–91.

———. "Dynamic Behaviour in Large Markets for Differentiated Products." *Review of Economic Studies*, April 1987, *54*(2), pp. 293–300.

McMillan, John. "Individual Incentives in the Supply of Public Inputs." *Journal of Public Economics*, August 1979, *12*(1), pp. 87–98.

Olson, Mancur. *The logic of collective action*. Cambridge, MA: Harvard University Press, 1965.

Pecorino, Paul. "Is There a Free-Rider Problem in Lobbying? Endogenous Tariffs, Trigger Strategies, and the Number of Firms." Unpublished manuscript, University of Alabama, December 1996.

Porter, Robert H. "Optimal Cartel Trigger-Price Strategies." *Journal of Economic Theory*, April 1983, *29*(2), pp. 313–38.

Potters, Jan and Sloof, Randolph. "Interest Groups: A Survey of Empirical Models That Try to Assess Their Influence." *European Journal of Political Economy*, November 1996, *12*(3), pp. 403–42.

Riezman, Raymond and Wilson, John D. "Politics and Trade Policy," in Jeffrey S. Banks and Eric A. Hanushek, eds., *Modern political economy*. Cambridge: Cambridge University Press, 1995, pp. 108–44.

Rodrik, Dani. "Tariffs, Subsidies, and Welfare with Endogenous Policy." *Journal of International Economics*, November 1986, *21*(3/4), pp. 285–99.

———. "Political Economy of Trade Policy," in Gene M. Grossman and Kenneth Rogoff, eds., *Handbook of international economics*, Vol. 3. Amsterdam: North-Holland, 1995, pp. 1457–94.

Sandler, Todd. *Collective action*. Ann Arbor, MI: University of Michigan Press, 1992.

Shapiro, Carl. "Theories of Oligopoly Behavior," in Richard Schmalensee and Robert D. Willig, eds., *Handbook of industrial organization*, Vol. 1. Amsterdam: North-Holland, 1989, pp. 329–414.

Deloitte & Touche LLP

Suite 2400
424 Church Street
SunTrust Center
Nashville, Tennessee 37219-2396

Independent Auditors' Report

Executive Committee
The American Economic Association
Nashville, Tennessee

We have audited the accompanying statements of financial position of The American Economic Association (the "Association") as of December 31, 1997 and 1996, and the related statements of unrestricted revenues, expenses and other changes in unrestricted net assets, changes in net assets and cash flows for the years then ended. These financial statements are the responsibility of the Association's management. Our responsibility is to express an opinion on these financial statements based on our audits.

We conducted our audits in accordance with generally accepted auditing standards. Those standards require that we plan and perform the audit to obtain reasonable assurance about whether the financial statements are free of material misstatement. An audit includes examining, on a test basis, evidence supporting the amounts and disclosures in the financial statements. An audit also includes assessing the accounting principles used and significant estimates made by management, as well as evaluating the overall financial statement presentation. We believe that our audits provide a reasonable basis for our opinion.

In our opinion, the financial statements referred to above present fairly, in all material respects, the financial position of The American Economic Association as of December 31, 1997 and 1996, and the changes in its net assets and its cash flows for the years then ended in conformity with generally accepted accounting principles.

Deloitte & Touche LLP

March 19, 1998

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF FINANCIAL POSITION
DECEMBER 31, 1997 AND 1996

	1997	1996
Assets		
CASH AND CASH EQUIVALENTS (Note 1)	$ 996,672	$1,189,540
INVESTMENTS (Notes 1 and 2)	9,109,988	7,589,942
ACCOUNTS RECEIVABLE, no allowance for doubtful accounts considered necessary	472,747	236,532
INVENTORY OF *Index of Economic Articles* (Note 1)	46,243	61,941
PREPAID EXPENSES	73,303	61,023
FURNITURE, FIXTURES AND EQUIPMENT, net (Note 1)	108,596	105,405
TOTAL ASSETS	$10,807,549	$9,244,383
Liabilities and Net Assets		
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	$ 810,154	$ 585,254
DEFERRED REVENUE: (Note 1)		
Membership dues and nonmember subscriptions	1,672,329	1,615,170
Job Openings for Economists	11,299	14,420
Total deferred revenue	1,683,628	1,629,590
ACCRUAL FOR SURVEY AND TELEPHONE DIRECTORY (Note 1)	61,085	124,550
Total liabilities	2,554,867	2,339,394
COMMITMENTS (Note 3)		
NET ASSETS:		
Unrestricted	8,238,254	6,877,459
Temporarily restricted (Note 1)	14,428	27,530
Total net assets	8,252,682	6,904,989
TOTAL LIABILITIES AND NET ASSETS	$10,807,549	$9,244,383

See notes to financial statements.

The American Economic Association Statements of Unrestricted Revenues, Expenses, and Other Changes in Unrestricted Net Assets Years Ended December 31, 1997 and 1996

	1997	1996
Unrestricted Operating Revenues:		
Membership dues and nonmember subscriptions (Note 1)	$2,106,468	$2,011,550
Job Openings for Economists subscriptions (Note 1)	23,029	33,535
Royalties (Note 5)	1,044,990	984,454
Advertising	146,743	132,564
Sale of *Index of Economic Articles*	16,182	75,515
Sale of copies, reprints and handbooks	41,516	49,279
Sale of mailing list	67,043	47,273
Annual meeting (net of associated costs of $369,185 and $329,761)	67,716	92,837
Investment return designated for current operations (Notes 1 and 2)	438,975	390,335
Other (Note 5)	78,526	79,159
Total unrestricted operating revenues	4,031,188	3,896,501
Net Assets Released from Restrictions:		
Satisfaction of program restrictions:	158,361	109,461
Total unrestricted operating revenues and other support	4,189,549	4,005,962
Operating Expenses:		
Publication:		
American Economic Review	925,492	939,611
Journal of Economic Literature	1,460,691	1,422,050
Journal of Economic Perspectives	500,996	496,789
Job Openings for Economists	79,172	77,769
Survey and *Telephone Directory* (Note 1)	126,813	39,919
Index of Economic Articles	120,103	174,210
	3,213,267	3,150,348
Management and general:		
General and administrative:		
Salaries and benefits (Note 4)	312,242	272,417
Rent (Note 3)	30,776	31,390
Other (Note 6)	286,270	277,291
Committees	135,273	124,054
	764,561	705,152
Total publication, management and general expenses (Note 7)	3,977,828	3,855,500
Program expenses	158,361	109,461
Total unrestricted operating expenses	4,136,189	3,964,961
Increase in Unrestricted Net Assets From Operations	53,360	41,001
Investment Return in Excess of Amounts Designated for Current Operations (Notes 1 and 2)	1,307,435	730,569
Increase in Unrestricted Net Assets	$1,360,795	$ 771,570

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN NET ASSETS
YEARS ENDED DECEMBER 31, 1997 AND 1996

	1997	1996
UNRESTRICTED NET ASSETS:		
Total unrestricted operating revenues	$4,031,188	$3,896,[illegible]
Net assets released from restrictions	158,361	109,[illegible]
Total unrestricted operating expenses	(4,136,1[illegible]	[illegible]
Investment return in excess of amounts designated for current operations	1,307,435	730,5[illegible]
Increase in unrestricted net assets	1,360,795	771,5[illegible]
TEMPORARILY RESTRICTED NET ASSETS:		
Grant proceeds:	145,259	88,[illegible]
Net assets released from restrictions	(158,361)	(109,[illegible]
Decrease in temporarily restricted net assets	(13,102)	(21,[illegible]
INCREASE IN NET ASSETS	1,347,693	750,[illegible]
NET ASSETS AT BEGINNING OF YEAR	6,904,989	6,154,[illegible]
NET ASSETS AT END OF YEAR	$8,252,682	$6,904,[illegible]

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CASH FLOWS
YEARS ENDED DECEMBER 31, 1997 AND 1996

	1997	1996
CASH FLOWS FROM OPERATING ACTIVITIES:		
Receipt of membership dues, nonmember subscriptions, royalties and other operating revenue	$3,410,036	$3,671,048
Disbursements to suppliers and employees	(3,772,362)	(3,764,434)
Grant proceeds	145,259	88,182
Grant disbursements	(158,361)	(109,461)
Proceeds from interest, dividends and capital gain distributions on investments	364,804	277,745
NET CASH (USED IN) PROVIDED BY OPERATING ACTIVITIES	**(10,624)**	**163,080**
CASH FLOWS FROM INVESTING ACTIVITIES:		
Purchases of investments	(680,471)	(2,076,739)
Proceeds from sale of investments	542,031	2,225,753
Purchases of furniture, fixtures and equipment	(43,804)	(33,457)
NET CASH (USED IN) PROVIDED BY INVESTING ACTIVITIES	**(182,244)**	**115,557**
NET (DECREASE) INCREASE IN CASH AND CASH EQUIVALENTS	**(192,868)**	**278,637**
CASH AND CASH EQUIVALENTS AT BEGINNING OF YEAR	1,189,540	910,903
CASH AND CASH EQUIVALENTS AT END OF YEAR	$ 996,672	$1,189,540
RECONCILIATION OF CHANGE IN NET ASSETS TO NET CASH USED IN OPERATING ACTIVITIES:		
Change in net assets	$1,347,693	$ 750,291
Adjustments to reconcile change in net assets to net cash (used in) provided by operating activities:		
Depreciation and amortization	40,613	35,270
Unrealized investment gains	(1,381,606)	(843,159)
Increase in accounts receivable	(236,215)	(18,140)
Decrease (increase) in inventory of *Index of Economic Articles*	15,698	(5,884)
Increase in prepaid expenses	(12,280)	(1,604)
Increase in accounts payable and accrued liabilities	224,900	11,697
Increase in deferred revenue	54,038	195,722
(Decrease) increase in accrual for *Survey* and *Telephone Directory*	(63,465)	38,887
NET CASH (USED IN) PROVIDED BY OPERATING ACTIVITIES	**$ (10,624)**	**$ 163,080**

See notes to financial statements.

The American Economic Association Notes to Financial Statements
Years Ended December 31, 1997 and 1996

1. Summary of Significant Accounting Policies

The American Economic Association (the "Association") is an educational organization whose purposes are [illegible] courage economic research, especially the historical and statistical study of the actual conditions of industrial [illegible] issue publications on economic subjects and to encourage freedom of economic discussion.

Cash and cash equivalents include highly liquid investments with original maturities of three months or less.

Investments are reported at fair value as reported by the respective funds. The Association designates only a portion of its cumulative investment return for support of current operations; the remainder is designated to support operations of future years and to offset potential market declines and other unforeseen contingencies. Investment return [illegible] senting five percent of total cash, cash equivalents and investments at the beginning of the year are designated [illegible] support current operations.

Inventory of Index of Economic Articles is stated at the lower of cost or market.

Furniture, fixtures and equipment is stated at cost net of accumulated depreciation. Depreciation is calculated [illegible] the straight-line method over the estimated useful lives of the assets.

Deferred revenue represents income from membership dues and nonmember subscriptions to the various periodicals of the Association which are deferred when received. These amounts are then recognized as operating revenue [illegible] ing the distribution of the specified publications to the members and subscribers of the Association. Income from [illegible] membership dues is recognized over the estimated average life of these members.

Accrual for Survey and Telephone Directory. Every four years, the Association publishes a *Survey* which lists, [illegible] other things, the names and addresses of its membership. It also publishes a *Telephone Directory* every four [illegible] which contains the name, address, electronic mail address and office or home phone and fax numbers of the [illegible] The *Survey* was published and distributed most recently in 1997 at no cost to the membership. The *Telephone Directory* was distributed most recently in 1995 at no cost to the membership. To properly match the publishing cost of [illegible] publications with revenue from membership dues, the Association provided approximately $47,000 and $40,000 [illegible] the years ended December 31, 1997 and 1996, respectively, for estimated publishing costs which will reduce [illegible] *Survey* and *Telephone Directory* expenses in the year of publication.

Temporarily restricted net assets represent amounts available for various training and student exchange programs [illegible] economic conferences and committees. Certain temporarily restricted net assets are administered on a reimbursement basis; therefore, disbursements are allowed prior to receipt of grant proceeds.

Federal income taxes. The Association files its federal income tax return as an educational organization substantially exempt from income tax under Section 501(c)(3) of the Internal Revenue Code. As required by Section 511(a) of [illegible] Code, the Association is subject to federal income taxes on certain revenues which are not substantially related [illegible] tax-exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing [illegible] The Association has been determined to be an organization which is not a private foundation.

Accounting estimates. The preparation of financial statements in conformity with generally accepted accounting [illegible] ciples requires management to make estimates and assumptions that affect the reported amounts of assets and [illegible] and disclosure of contingent assets and liabilities at the date of the financial statements. Estimates also [illegible] reported amounts of revenue and expenses during the reporting period. Actual results could differ from those [illegible]

Certain reclassifications have been made to the 1996 financial statements to conform to the 1997 presentation.

Investments and Investment Return

Investments consist of:

	1997	1996
Money market funds	$ 315	$ 300
Mutual funds:		
Equity	7,176,187	5,826,447
Debt	1,933,486	1,763,195
	9,109,673	7,589,642
	$9,109,988	$7,589,942

Investment return consists of:

	Years Ended December 31,	
	1997	1996
Dividends and unrealized gains and losses, net	$1,660,430	$ 993,462
Net realized gain on sales	60,513	104,298
Interest	25,467	23,144
	1,746,410	1,120,904
Investment return designated for current operations	(438,975)	(390,335)
Investment return in excess of amounts designated for current operations	$1,307,435	$ 730,569

3. Commitments

The Association leases office space under cancelable and noncancelable operating leases. Total rental expense under these leases was approximately $88,000 and $83,000 during the years ended December 31, 1997 and 1996, respectively.

The minimum future rental commitments under noncancelable operating leases at December 31, 1997 are as follows:

Year ending December 31,	
1998	$ 90,256
1999	84,870
2000	45,129
2001	3,970
	$224,225

The Association has contracts for hotel rooms and facilities for various meetings through 2001.

4. Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employees' compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was approximately $82,000 and $78,000 for the years ended December 31, 1997 and 1996, respectively.

5. Royalties and Other Unrestricted Operating Revenues

Royalties for the years ended December 31, 1997 and 1996, consist of the following:

	1997	1996
SilverPlatter	$848,653	$798,451
OCLC	84,257	74,830
EconLit	57,270	50,700
Dialog/Knight-Ridder	32,403	34,616
Miscellaneous	22,407	25,857
	$1,044,990	**$984,454**

Other unrestricted operating revenues for the years ended December 31, 1997 and 1996, consist of the following

	1997	1996
AER submission fees	$ 57,731	$ 55,050
CSWEP membership dues	20,544	24,109
Other donations	251	—
	$ 78,526	$ 79,159

6. Other General and Administrative Expenses

Other general and administrative expenses for the years ended December 31, 1997 and 1996, consist of the follo

	1997	1996
Dues and subscriptions	$ 77,936	$ 77,786
Accounting and legal	31,127	36,106
Mailing list file maintenance	32,100	31,110
Bank charges	30,509	27,590
Postage	25,903	24,895
Depreciation	29,402	24,368
Election expenses	16,144	15,202
Insurance and miscellaneous	14,719	13,017
Office supplies	9,442	10,836
Telephone	7,447	6,856
Investment counsel and custodian fees	2,412	6,277
President and president-elect expenses	9,129	3,248
	$286,270	**$277,291**

7. Ratio of Unrestricted Net Assets to Expenses

The ratio of unrestricted net assets at December 31, 1997, to 1998 budgeted publication, management and g expenses is 1.82. The ratio of unrestricted net assets at December 31, 1996, to 1997 actual publication, manag and general expenses is 1.73.